우리말 자연어 처리 기술
- 과거와 현재

김학수
강원대학교 컴퓨터정보통신공학전공 교수

1. 서론

최근 인공 지능(AI: Artificial Intelligence) 열풍과 함께 애플의 시리(Siri), 아마존의 알렉사(Alexa), 케이티의 기가지니(GiGAGenie), 네이버의 클로바(Clova) 등 개인 비서 서비스가 활발히 개발되고 있다. 이러한 인공 지능 개인 비서 서비스의 핵심은 사용자의 말을 얼마나 잘 알아듣고 똑똑한 대답을 내놓느냐 하는 것이다. 예를 들어, 사용자가 "너무 더운 거 같지 않니?"라고 했을 때, 똑똑한 인공 지능 개인 비서라면 "에어컨을 가동할까요?"라고 대답을 할 것이다. 인간의 기준에서 보면 "에어컨을 가동할까요?"라고 대답하는 것이 대수롭지 않겠지만 기계[1]에 그것을 이해시키기 위해서는 몇 단계의 자연어 처리 과정이 필요하다. 먼저, 입력된 문장이 어떤 단어들로 구성되어 있는지를 분석해야 한다. 즉, 입력 문장이 '너무(부사) 덥-(형용사)+-ㄴ(어미) 거(의존 명사) 같-(형용사)+-지(어미) 않-(보조 용언)+-니(어미)+?(문장 부호)'와 같이 구성되어 있음을 밝혀야 한다. 다음으로는 문장의 주어와 목적어가 무엇인지를 찾아야 한다. 예를 든 문장의

1) 이 글에서 '기계'라고 함은 인공 지능 소프트웨어(Software)를 의미한다.

경우에 '너는'이라는 주어가 생략되어 있다는 것도 밝혀야 한다. 다음으로 '덥'이라는 단어가 '기온이 높다'라는 의미가 있다는 것을 밝혀야 한다. 마지막으로 입력된 문장에 숨겨진 의도가 긍정·부정 질문(Yes/No Question)이 아니라 요구(Request)라는 것을 찾아내야 한다. 만약 해당 문장을 긍정이나 부정 질문으로 이해하게 되면 "예, 덥습니다."라는 엉뚱한 답변을 하게 될 것이다. 예를 든 문장에서 살펴본 것과 같이 인간의 기준에서 단순한 문장이더라도 기계를 이해시키기 위해서는 <표 1>과 같이 여러 단계에 걸친 복잡한 자연어 처리 과정이 필요하다.

표 1 자연어 처리 단계

단계	설명	보기: 나는 그 과자를 먹었다.
형태소 분석	문장을 형태소 열로 분리하고 품사를 부착하는 단계	나(대명사)+는(조사) 그(대명사) 과자(명사)+를(조사) 먹-(동사)+-었-(선어말 어미)+-다(어말 어미)+.(문장 부호)
구문 분석	문장의 문법적 적합성과 어절의 구문적 역할(주어, 목적어 등)을 찾는 단계	[SUBJ: 나는 [[MOD: 그 [OBJ: 과자를]] 먹었다]]
의미 분석	문장을 구성하는 술어와 인자들 사이의 의미적 적합성을 분석하는 단계	PREDICATE: 먹다 AGENT: 나/ANIMATE OBJECT: 그 과자/EATABLE
담화 분석	대화 문맥을 파악하여 상호 참조를 해결하고 의도를 파악하는 단계	SPEECH ACT: STATEMENT PREDICATE: 먹다 AGENT: 홍길동/ANIMATE OBJECT: 과자/EATABLE

이 글에서는 우리말 자연어 처리를 위해 전통적으로 사용된 방법들과 문제점들을 살펴본다. 그리고 기계 학습의 발달로 인해 기존 자연어 처리 문제점들이 어떻게 극복되고 있는지 알아보고 향후 발전 방향을 제시한다. 이 글의 목적은 자연어 처리에 대한 전문 지식이 없는 일반 독자들에게

가능한 한 쉬운 용어와 예제로 우리말 자연어 처리 기술에 대한 보편적 이해를 돕고자 하는 것임을 미리 밝힌다.

2. 전통적 자연어 처리 기술

2.1. 알고리즘 기반 후보 생성

전통적 자연어 처리 기술들은 알고리즘을 이용하여 여러 개의 후보를 생성하고, 확률적인 방법으로 애매성(ambiguity)을 해소하는 접근법을 사용하였다. 형태소 분석에 가장 많이 사용되어 온 알고리즘은 태뷸러 파싱(Tabular Parsing) 알고리즘이다. <그림 1>은 '감기는'이라는 어절을 자소 단위로 분리한 후, 오른쪽(마지막 자소)부터 왼쪽으로 이동하면서 형태소 분석을 수행하는 태뷸러 파싱 알고리즘의 적용 예를 보여 준다.

그림 1 오른쪽 우선 태뷸러 파싱 알고리즘을 이용한 형태소 분석

<그림 1>에서 보듯이 형태소 분석을 위해서는 사전(형태소 어휘와 품사를 저장하고 있는 자료 구조)과 접속 정보(형태소 어휘 또는 품사의 연속 가능 여부를 담고 있는 자료 구조)가 필요하다. 태뷸러 파싱 알고리즘은 자소를 결합하여 형태소 후보를 생성하고 사전과 접속 정보를 이용하여 조합 가능한 모든 형태소 열을 생성한다. <그림 1>의 경우에 오른쪽 상단에 '감기는'이라는 어절로부터 모든 가능한 형태소 후보 열인 '감기(명사)+는(보조사), 감(동사)+-기(명사형 전성 어미)+는(보조사), 가-(동사)+-ㅁ(명사형 전성 어미)-+기(명사형 전성 어미)+는(보조사)'이 생성된 것을 볼 수 있다.

의존 구조 분석도 형태소 분석과 비슷한 접근 방법을 사용한다. 어절과 어절 사이의 관계를 의존 문법으로 기술하고 해당 문법을 바탕으로 알고리즘을 적용하여 모든 가능한 의존 구조 분석 트리(tree)를 생성한다. <표 2>는 의존 문법의 일부를 이해하기 쉽도록 표로 정리한 것이다.

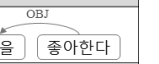
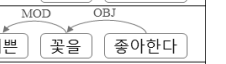
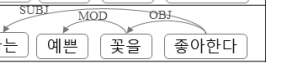
표 2 의존 문법 예시

관계	의존소의 대표 품사	지배소의 대표 품사
수식	관형사, 관형격 조사, 관형사형 전성 어미, 명사, 부사	명사
수식	관형사, 관형격 조사, 관형사형 전성 어미, 부사	대명사
부가	주격 조사, 목적격 조사, 부사격 조사, 보조사, 부사, 연결형 서술 어미, 부사형 전성 어미	동사, 형용사
강조	부사	부사, 관형사

<표 2>에서 보듯이 입력된 문장의 각 어절은 대표 품사 형태로 변환된다. 그리고 지배 가능한 모든 어절들을 확인하면서 의존 구조 트리 후보들을 생성한다. <그림 2>는 "나는 예쁜 꽃을 좋아한다."라는 문장의 가장 오른쪽 어절인 '좋아한다'부터 첫 어절인 '나는'까지 지배 가능 경로를 따라 이동하

면서 의존 구조 트리 후보들을 생성하는 과정을 보여 준다(김창현 외, 1993).

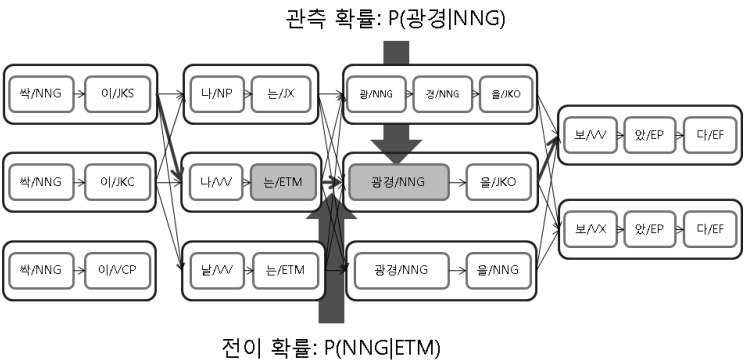
그림 2 지배 가능 경로를 이용한 오른쪽 우선 의존 구조 분석의 예

단계	의존소 후보 어절	지배소 후보 어절	의존문법 검사	의존구조 트리
1	꽃을	좋아한다	OBJ	
2	예쁜	꽃을	MOD	
3	예쁜	좋아한다	X	
4	나는	예쁜	SUB	
5	나는	좋아한다	SUB	

2.2. 확률 기반 애매성 해소

규칙이나 알고리즘에 기반한 자연어 처리 모델(Model)은 자연어 본연의 특성으로 인해 다수의 분석 후보들을 생성한다. <그림 2>만 보더라도 '나는'이 '예쁜'을 꾸미는 경우와 '좋아한다'를 꾸미는 경우가 생성된다. 그러므로 문맥을 바탕으로 어떤 것이 보편타당한 것인지를 선택하는 애매성 해소 과정이 필요하다. 전통적인 애매성 해소 방법은 대용량의 언어 지식 부착 말뭉치로부터 통계 데이터를 습득하고 이를 바탕으로 확률적 선택을 하는 것이다. 형태소 분석의 경우에 품사 정보가 부착된 대용량의 말뭉치로부터 해당 품사에서 형태소 후보가 관측될 확률과 현재 품사가 다음 품사로 전이될 확률의 곱을 계산한 후, 가장 높은 확률값을 가지는 형태소 열을 선택하는 HMM(Hidden Markov Model)을 주로 사용한다(Charniak 외, 1993; Lee & Rim, 2005). <그림 3>은 HMM에 기반하여 "싹이 나는 광경을 보았다."라는 문장으로부터 최적의 품사 부착 형태소 열을 찾아내는 과정을 보여 준다.

그림 3 HMM을 이용한 형태소 분석(품사 부착)의 예



의존 구조 분석도 형태소 분석과 비슷한 방법으로 애매성을 해소한다. 식 (1)에서 보는 것과 같이 의존 구조 정보가 부착된 대용량의 말뭉치로부터 두 어절이 특정 구문 관계로 출현할 확률을 계산하고 최댓값을 갖는 의존 구조 트리를 선택하는 방법을 주로 사용한다(김학수 외, 1997; Nivre, 2005).

식 (1) $$P(T|S) \approx \prod_{i=1}^{n} P(D|E_i, E_{Head(i)})$$

식 (1)에서 S와 T는 입력 문장과 의존 구조 트리를 의미하며, E_i와 $E_{Head(i)}$는 문장 내 i번째 어절과 그 어절의 지배소 후보가 되는 어절을 의미한다. 그리고 D는 E_i와 $E_{Head(i)}$ 사이의 구문 표지를 의미한다.

3. 최신 자연어 처리 기술

3.1. 기계 학습의 도입

최근 딥러닝(deep learning)을 필두로 기계 학습 기술들이 눈부시게 발달함에 따라 자연어 처리 분야에서도 기계 학습 모델을 이용하여 애매성을 해소하고자 하는 연구들이 활발히 진행되고 있다. 위키백과(Wikipedia)에 정의되어 있는 기계 학습의 뜻을 살펴보면 다음과 같다.

기계 학습(機械學習) 또는 머신 러닝(machine learning)은 인공 지능의 한 분야로, 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야를 말한다. 기계 학습의 핵심은 표현(representation)과 일반화(generalization)에 있다. 표현이란 데이터의 평가이며, 일반화란 아직 알 수 없는 데이터에 대한 처리이다.

즉, 기계 학습이란 주어진 데이터(학습 데이터라고 불림)의 특성을 잘 설명할 수 있으면서 새로운 데이터(평가 데이터라고 불림)에 적응성이 높은 일종의 함수를 자동으로 찾아내는 것이라고 할 수 있다.

자연어 처리와 연관된 기계 학습 문제는 분류 문제(classification problem), 순차적 표지 부착 문제(sequence labeling problem), 그리고 정책 결정 문제(policy decision problem)로 나뉠 수 있다. <표 3>은 각 기계 학습 문제와 관련이 있는 자연어 처리 기술들을 보여 준다.

표 3 기계 학습 문제에 따른 자연어 처리 기술[2)]

기계 학습 문제	자연어 처리 기술	설명
분류 문제	의존 구조 분석	두 어절 사이의 의존 여부와 구문 표지를 결정
순차적 표지 부착 문제	형태소 분석	문장을 구성하는 어절 열을 형태소 열로 분할하고 품사를 부착
	개체명 인식	문장 내에 존재하는 개체명을 탐지하고 개체명 범주를 부착
	의미역 결정	문장 내에 존재하는 서술어를 탐지하고 서술어를 수식하는 어절에 의미역 범주를 부착
정책 결정 문제	대화 상태 추적	대화를 구성하는 현재 발화에 대한 다음 발화의 의도를 결정

분류 문제는 입력 데이터가 어떤 범주에 속하는지를 결정하는 문제이다. 분류 문제가 적용된 예로는 스팸 메일 자동 필터링 시스템, 신문 기사 범주 자동 결정 시스템, 댓글 감정 분석 시스템 등이 있다. 자연어 처리 분야에서는 의존 구조 분석 시에 현재 어절의 지배소가 어느 것인지를 결정하는 것(후속 어절 각각에 대해 지배소 가능 여부를 결정하는 것)에 적용될 수 있다. 순차적 표지 부착 문제는 시간 순서대로 입력되는 데이터 전체를 고려하면서 각 지점에 분류 모델을 적용하여 범주를 부착하는 것이다. 어절로 구성된 문장 전체에 품사를 부착하는 형태소 분석이나 개체명 경계를 찾고 범주를 부착하는 개체명 분석 등 자연어 처리의 많은 문제들이 순차적 표지 부착 문제에 해당된다. 정책 결정 문제는 대량의 데이터를 바탕으로 정해진 목적을 달성하기에 가장 적합한 정책이 무엇인지를 결정하는 문제이다. 구글(Google)의 딥마인드(DeepMind)가 선보인 '알파고(AlphaGo)'와 같은 프로그램이 정책 결정 문제가 적용된 예라고 할 수 있다. 자연어 처리 분야에서

2) <표 3>의 분류는 편의상 구분한 것이며, 각 자연어 처리 기술들이 해당 기계 학습 문제에 반드시 종속됨을 의미하는 것은 아니다.

는 대용량의 목적 지향 대화 말뭉치(Goal-Oriented Dialogue Corpus)를 바탕으로 해당 목적을 달성하기 위해서 현재 어떤 의도의 발화를 하는 것이 가장 좋은 정책인지를 결정하는 것에 적용될 수 있다.

3.2. 기계 학습 기반 자연어 처리 모델

기계 학습 기반의 형태소 분석, 개체명 인식, 의미역 결정 등은 식 (2)와 같이 순차적 표지 부착 문제로 변환하여 해결한다.

식 (2) $$NLP(S) = \arg_{L_{1,n}} \max P(L_{1,n} | Seg_{1,n})$$

식 (2)에서 S는 입력된 문장을, $Seg_{1,n}$은 표지를 부착해야 하는 n개의 분절 열을, $L_{1,n}$은 $Seg_{1,n}$에 부착되는 n개의 표지 열을 의미한다. 일반적으로 형태소 분석을 위한 분절은 음절이 되며, 개체명 인식을 위한 분절은 형태소가 되고, 의미역 결정을 위한 분절은 각 서술어를 수식하는 어절이 된다. 예를 들어, “청와대에 갔다.”라는 문장을 형태소로 분석하기 위한 분절 열은 공백을 포함하여 ‘$Seg_{1,8}$=[청, 와, 대, 에, _, 갔, 다, .]’가 되며, 개체명 인식을 위한 분절 열은 ‘$Seg_{1,7}$=[청와대, 에, _, 가, 았, 다, .]’가 된다. 의미역 결정을 위한 분절 열은 ‘$Seg_{1,2}$=[청와대에, 갔다.]’가 된다. 형태소 분석과 개체명 인식을 위한 표지는 <표 4>와 같이 경계를 나타내는 표지와 범주를 나타내는 표지를 결합하여 사용한다.

표 4 형태소 분석, 개체명 인식, 의미 분석을 위한 표지

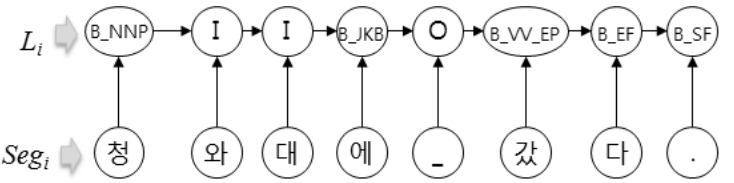
구분	표지	설명	예제
형태소 분석	B_POS	B: 형태소 경계 시작 POS: 품사	$Seg_{1,8}$=[청, 와, 대, 에, _, 갔, 다, .] $L_{1,8}$=[B_NNP, I, I, B_JKB, O, B_VV_EP, B_EF, B_SF] NNP: 고유명사, JKB: 부사격 조사, VV: 동사, EP: 선어말 어미, EF: 어말 어미, SF: 마침 기호
	I	I: 형태소 경계 내부	
	O	O: 형태소 경계 외부	
개체명 인식	B_NEC	B: 개체명 경계 시작 NEC: 개체명 범주	$Seg_{1,7}$=[청와대, 에, _, 가, 았, 다, .] $L_{1,7}$=[B_LOC, O, O, O, O, O, O] LOC: 장소
	I	I: 개체명 경계 내부	
	O	O: 개체명 경계 외부	
의미역 결정	B_SR	B: 의미역 경계 시작 SR: 의미역 범주	$Seg_{1,4}$=[나는, 혼자, 청와대에, 갔다.] $L_{1,4}$=[B_AGT, O, B_LOC, B_PRED] AGT: 행위주역, LOC: 장소역, PRED: 술어
	I	I: 의미역 내부	
	O	O: 비의미역	

식 (2)를 단순화하려면 현재 표지는 현재 분절의 특성에만 영향을 받는다는 독립 가정을 적용하고, 현재 표지는 바로 이전 표지에 영향을 받는다는 1차 마코프(Markov) 가정을 적용하면 식 (3)과 같다.

식 (3) $$NLP(S) = \arg_{L_{1,n}} \max \prod_{i=1}^{n} P(L_i|Seg_i)P(L_i|L_{i-1})$$

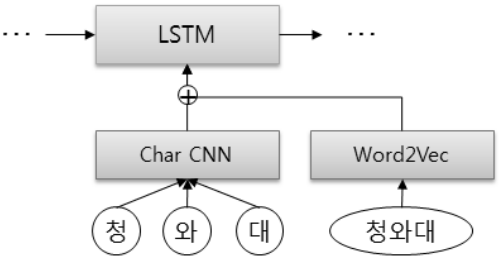
식 (3)을 개념적으로 표현하면 <그림 4>와 같다.

그림 4 순차 표지 부착 기반 자연어 처리 모델의 개념도



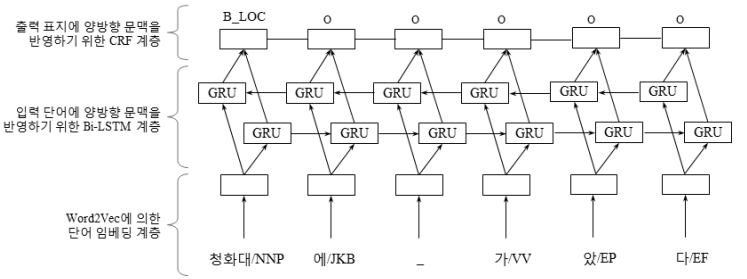
식 (3)과 같은 모델에서 성능을 좌우하는 것은 Seg_i를 어떻게 표현할 것인지(어떻게 추상화할 것인지)와 L_i 주변의 문맥을 어떻게 반영할 것인지로 요약할 수 있다. MEMM(Maximum Entropy Makov Model)(Reynar & Ratnaparkhi, 1997), CRFs(Conditional Random Fields)(Lafferty 외, 2001)와 같은 전통적 기계 학습 모델들에서는 Seg_i를 효과적으로 추상화하기 위해서 언어 분석 전문가들이 정의한 자질들(features)을 사용하였다. 예를 들어, 개체명 인식의 경우에 Seg_i의 전체나 일부가 개체명 사전에 포함되어 있는지 여부, Seg_i의 품사, Seg_{i-1}의 품사, Seg_{i+1}의 품사 등이 주요 자질로 사용되었다. 그러나 최근에 딥뉴럴넷(Deep Neural Network)에 대한 연구가 활발히 이루어지면서 대용량의 말뭉치에서 자동으로 계산된 Word2Vec(Mikolov 외, 2013)을 이용하여 Seg_i를 벡터 형태로 추상화하거나 Seg_i를 음절 형태로 분리한 후 CNN(Convolutional Neural Network)(Krizhevsky 외, 2012)을 통해 추상화하는 방법이 사용된다. <그림 5>는 Seg_i를 Word2Vec과 CNN을 이용하여 추상화하는 개념도이다.

그림 5 딥뉴럴넷을 이용한 Seg_i 추상화 개념도



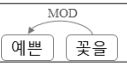
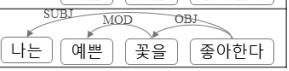
식 (3)의 경우에 현재 표지의 왼쪽에 있는 표지만 문맥으로 고려되는 단방향성 문제가 존재한다. 이를 해결하고 효과적으로 L_i 주변 문맥을 반영하기 위해서 좌우 문맥을 모두 고려하는 CRFs 모델이 제안되었으며, 최근에는 RNN(Recurrent Neural Network) 계열의 딥뉴럴넷과 CRFs가 결합된 모델들이 주로 사용된다. <그림 6>은 좌우 문맥을 모두 고려하는 Bi-LSTM-CRF(Bidirectional Long Short-Term Memory with Conditional Random Fields)라는 신경망 모델의 개념도이다(Huang 외, 2015).

그림 6 개체명 인식을 위한 Bi-LSTM-CRF 모델의 개념도



기계 학습 기반의 의존 구조 분석은 두 어절 사이의 의존 여부와 구문 표지를 결정적으로 찾아가는 분류 문제로 변환하여 해결할 수 있다. <그림 7>은 전이 기반 구문 분석 알고리즘에 따라 "나는 예쁜 꽃을 좋아한다."라는 문장의 의존 구조를 분석하는 과정을 보여 준다(Sagae & Tsujii, 2008).

그림 7 전이 기반 한국어 의존 구조 분석의 예

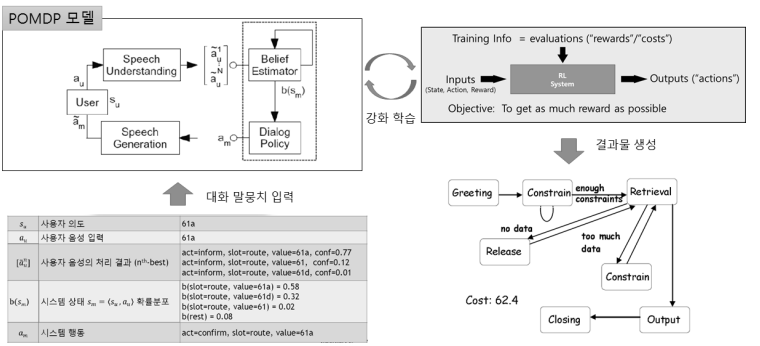
단계	스택(Stack)	큐(Queue)	분류	의존구조 트리
1	나는	예쁜 꽃을 좋아한다	Shift	
2	나는 예쁜	꽃을 좋아한다	Reduce	
3	나는	꽃을 좋아한다	Shift	
4	나는 꽃을	좋아한다	Reduce	
5	나는	좋아한다	Reduce	

<그림 7>에서 보듯이 전이 기반 한국어 의존 분석은 스택(후입 선출 형태의 자료 구조)과 큐(선입 선출 형태의 자료 구조)의 첫 어절을 비교하여 'Shift'할 것인지 'Reduce'할 것인지를 반복적으로 이진 분류(binary classification) 하는 방식으로 이루어진다. 'Shift'인 경우에 큐의 첫 어절을 스택으로 이동시키며, 'Reduce'인 경우에 스택의 마지막 어절과 큐의 첫 어절로 구성된 의존 구조 트리를 구성하고 스택의 마지막 어절을 제거한다. 이와 같은 전이 기반 방법은 두 어절 사이의 정보를 바탕으로 의존 여부를 결정하기 때문에 문장의 전체 구조를 반영할 수 없다는 단점이 있다. 이러한 문제를 해결하기 위해서 최근에는 모든 의존 관계 후보를 그래프(graph) 형태로 만들고 전체 후보 중에서 가장 높은 점수를 갖는 후보를 효과적으로

선택하기 위한 다양한 방법들이 활발히 연구되고 있다.

기계 학습 기반의 대화 상태 추적은 강화 학습으로 대변되는 정책 결정 문제로 변환하여 해결한다. 대화 상태 추적에 적용된 대표적 정책 결정 모델은 POMDP(Partially Observable Markov Decision Process) 모델이다. POMDP 모델은 대용량의 대화 말뭉치로부터 강화 학습(Reinforcement Learning)을 통해 시스템의 대화 상태를 확률적으로 파악하고 다음 행동(Action)을 결정한다. <그림 8>은 POMDP 기반 대화 상태 추적 과정을 개념적으로 표현한 것이다(Kim 외, 2013; Jang 외, 2016; 김학수, 2017).

그림 8 POMDP 기반 대화 상태 추적 개념도

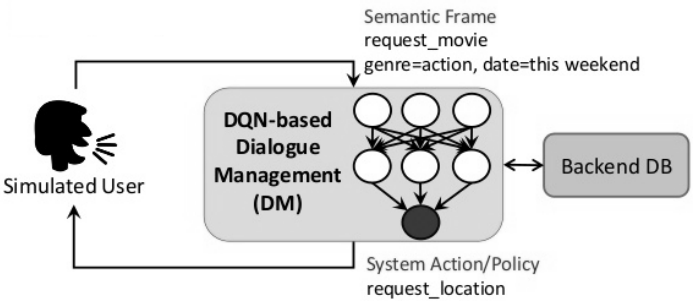


s_u	사용자 의도	61a
a_u	사용자 음성 입력	61a
$[\tilde{a}_u^N]$	사용자 음성의 처리 결과 (n^{th}-best)	act=inform, slot=route, value=61a, conf=0.77 act=inform, slot=route, value=61, conf=0.12 act=inform, slot=route, value=61d, conf=0.01
$b(s_m)$	시스템 상태 $s_m = (s_u, a_u)$ 확률분포	b(slot=route, value=61a) = 0.58 b(slot=route, value=61d) = 0.32 b(slot=route, value=61) = 0.02 b(rest) = 0.08
a_m	시스템 행동	act=confirm, slot=route, value=61a

<그림 8>과 같은 확률 기반 모델은 음성 인식 오류를 포함한 예기치 않은 사용자 입력에 대해서 부분적으로 관찰된 정보를 바탕으로 시스템의 의도를 결정할 수 있다는 장점이 있지만 학습을 위해서 매우 많은 담화 정보 부착 말뭉치를 필요로 한다는 문제를 안고 있다. 예를 들어, <그림 8>의 결과물로 제시된 상태 전이 모델은 약 71만 개의 항공 예약 관련 대화 말뭉치로부터 학습된 것이다. 최근에는 딥뉴럴넷을 이용하여 다음 행동을 결정하는 NNMDP(Neural Network Markov Decision Process)와 같은

모델들이 연구되고 있다. <그림 9>는 대화 상태 추적을 위해 Deep Q-Network가 사용되는 개념도이다(Li, 2017).

그림 9 Deep Q-Network을 이용한 대화 상태 추적 개념도



4. 결론

이 글에서는 우리말 자연어 처리를 위한 전통적 모델부터 최신 기계 학습 모델에 이르기까지 어떤 변화가 있었으며 풀고자 하는 문제가 어떤 것이었는지 개념적으로 설명하였다. 먼저, 알고리즘 기반의 전통적 형태소 분석법과 구문 분석법을 소개하고, 애매성 해소를 위한 확률적 접근 방법을 설명하였다. 그리고 기계 학습이 적용된 최신 자연어 처리 모델들을 분류 문제, 순차적 표지 부착 문제, 정책 결정 문제로 묶고, 각 기계 학습 문제가 자연어 처리 기술에 어떻게 적용되는지 설명하였다. 지금까지 살펴본 연구 흐름을 토대로 향후 자연어 처리 기술의 발전 방향은 다음과 같을 것으로 생각된다.

그동안 기계 학습 기반 자연어 처리 기술의 가장 큰 걸림돌은 성능 향상에 기여하는 좋은 자질(Feature)을 어떻게 선별하느냐 하는 것이었다. 그런데

딥뉴럴넷을 이용하면 좋은 자질의 선별이 자동으로 이루어지기 때문에 누구나 빠르고 쉽게 자연어 처리 응용 시스템을 개발할 수 있게 되었다. 이러한 편의성 때문에 딥뉴럴넷을 활용한 자연어 처리 연구가 당분간 대세를 이룰 것이라는 것은 의심의 여지가 없어 보인다. 그러나 딥뉴럴넷에 의한 자질의 선별은 자연어 처리 전문가의 입장에서 보면 아직 아쉬운 부분이 많이 존재한다. 그러므로 언어 처리 전문가들에 의해 선별된 자질과 딥뉴럴넷에 의해 자동 선별된 자질을 효과적으로 결합하여 성능을 향상시키고자 하는 시도가 꾸준히 진행될 것으로 보인다. 마지막으로 현재 파이프라인(pipeline) 형태로 구성된 자연어 처리 기술들은 이전 단계의 오류가 다음 단계로 전파되는 문제가 존재한다. 즉, 형태소 분석에 오류가 포함되면 이후 단계인 구문 분석, 의미 분석, 담화 분석에 오류가 전파되면서 성능이 급격히 떨어지게 된다. 이러한 문제를 해결하려고 각 자연어 처리 기술들을 유기적으로 결합하는 통합 모델에 대한 연구가 이루어질 것으로 예상된다.

참고 문헌

김창현, 김재훈, 서정연(1993), “지배 가능 경로를 이용한 오른쪽 우선 구문 분석”, ≪제5회 한글 및 한국어정보처리 학술대회 자료집≫, 35-44쪽.

김학수, 김지훈, 서정연(1997), “통계적 처리방법을 이용한 한국어 의존 구조 분석기”, ≪1997년도 인지과학회 춘계학술발표 논문집≫, 200-209쪽.

김학수(2017), “인공 지능 음성언어 비서 시스템의 자연언어처리 기술들”, ≪정보과학회지≫ 35-8, 9-18쪽.

E. Charniak, C. Hendrickson, N. Jacobson, and M. Perkowitz(1993) “Equations for part-of-speech tagging,” in Proceedings of the 11th National Conference on Artificial Intelligence, pp. 784 - 789.

Z, Huang, W. Xu, and K. Yu(2015), Bidirectional LSTM-CRF Models for Sequence Tagging, arXiv:1508.01991.

Y. Jang, J. Ham, B.-J. Lee, Y. Chang, and K.-E. Kim(2016), “Neural Dialog State Tracker for Large Ontologies by Attention Mechanism”, Proceedings of IEEE Workshop on Spoken Language Technology.

D. Kim, J. Choi, K.-E. Kim, J. Lee, and J. Sohn(2013), “Engineering Statistical Dialog State Trackers: A Case Study on DSTC”, Proceedings of the SIGDIAL 2013 Conference, pp. 462-466.

A. Krizhevsky, I. Sutskever, and G. E. Hinton(2012), “Imagenet classification with deep convolutional neural networks,” Proceedings of Conference on Neural Information Processing Systems.

J. Lafferty, A. McCallum, and F. Pereira(2001), “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” Proceedings of the International Conference on Machine Learning, pp. 282 - 289.

D. Lee and H. Rim(2005), “Probabilistic models for Korean morphological analysis,” Proceedings of the International Joint Conference on Natural Language Processing, pp. 197 - 202.

Y. Li(2017), Deep reinforcement learning: An overview, arXiv:1701.07274.

T. Mikolov, K. Chen, G. Corrado, and J. Dean(2013), “Efficient estimation of word representations in vector space,” Proceedings of International Conference on Learning Representation(arXiv: 1301.3781).

J. Nivre(2005), Dependency grammar and dependency parsing, MSI Technical Report 05133.

J. C. Reynar, A. Ratnaparkhi(1997), “A maximum entropy approach to identifying sentence boundaries,” Proceedings of Fifth Conference on Applied Natural Language Processing, pp. 16 - 19.

K. Sagae and J. Tsujii, “Shift-Reduce Dependency DAG Parsing,” Proceedings of the 22nd International Conference on Computational Linguistics, pp. 753 - 760.