creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학박사학위논문

강화 학습을 이용한 단어 감정 값 및 진술문 상태 값 측정법 연구

2018년 8월

서울대학교 대학원

협동과정 인지과학 전공

김 영 삼

국문 초록

본 논문에서 강화 학습(reinforcement learning)이란 기계 학습 방법의 한 종류로 환경으로부터 주어지는 보상을 최대화하는 방식의 학습 모형을 가리킨다. 이 방법론이 기계 학습 방법의 주류인 지도 학습(supervised learning)과 원리적으로 다른 점은 이 방법론이 에이전트가 무엇을 할 지에 대한 구체적인 지도(supervision)를 하지 않고, 에이전트가 처한 환경 안에서 여러 시행을 통해 그것을 발견하도록 한다는 점에 있다. 이 과정에서 에이전트가 취하는 행위, 즉 행동은 바로 주어지는 보상 외에도 계속적으로 이어지는 상황들에 따른 보상들을 최대화하는 것을 목적으로 하게 된다.

본 연구의 주요 목적은 이 강화 학습 방법론을 통해 자연어 텍스트를 점진적(incremental), 연속적(continuous)으로 처리하는 방법을 연구하는 데 있다. 자연어 처리, 혹은 전산 언어학은 다양한 과제를 가지고 있는 방대한 분야이므로 본 논문에서는 크게 두 가지 종류의 과제, 상태 값 측정 과제와 상태 유형 예측 과제를 중점적으로 다룬다. 전자의 경우에는 어휘의 감정 극성 값을 강화 학습의 핵심적 알고리즘인 시간차(temporal difference) 알고리즘을 통해 측정하는 방법을 연구하고, 후자의 경우에는 의료 텍스트 문장의 연속적 상태 측정을 강화 학습 방법론을 통해 수행하는 방법을 살펴볼 것이다.

자연어 텍스트의 점진적, 혹은 연속적 처리 과정은 인간의 언어 심리학적 처리 양상을 볼 때 인지 친화적인 접근법으로 보이는데, 왜냐하면 여러 심리학적, 신경학적 연구결과들을 통해 언어 처리의 과정이 기본적으로 자동적, 점진적 처리임이 알려져 있기 때문이다. 본 논문에는 그런 측면에서 강화 학습을 보다 인지 모형에 입각하여 활용하고자 하는 노력이 들어 있다. 마지막 논의 부분에서는 인지모형 기반 강화 학습에 대해 논의하고 이 방법을 자연어 처리에 어떻게 활용할 것인지에

대한 전망을 다루었다.

주요어: 강화학습, 자연어처리, 단어 감정값, 진술문 상태값, 점증적 처리
학번: 2009-30810

차 례

국문 초록 **iii**

그림 차례 **ix**

표 차례 **xii**

1 서론 **1**

1.1 강화 학습의 짧은 역사 2

1.2 연구 질문들 4

1.3 논문의 전체적 구성 5

I 연구 배경 **7**

2 언어 처리의 점진성 **9**

2.1 통사적 처리의 점진성 9

2.2 의미적 처리의 점진성 10

3 강화 학습 **13**

3.1 강화 학습 모형 14

3.2 마르코프 결정 과정 14

3.3 목적, 상태, 행위 그리고 보상 16

3.4 강화 학습 알고리즘의 두 특징 17

4 가치 예측 문제에 대한 시간차 학습 **19**

4.1 마르코프 보상 과정 19

4.2 가치 함수 20

4.3 몬테 카를로 방법 20

4.4 시간차 예측 21

5 모형 기반 강화 학습과 함수 근사법 **25**

5.1 모형과 계획 25

5.2 함수 근사와 상태 표상 28

5.2.1 모수적 방법 29

5.2.2 비모수적 방법 30

II 실험 **33**

6 시간차 기반 단어 감정 극성 값 측정법 **35**

6.1 실험의 기본적 설계 35

6.1.1 상태 36

6.1.2 보상 36

6.2 실험 1 37

6.2.1 데이터 37

6.2.2 비교 조건 37

6.2.3 실험 방법 39

6.2.4 결과 41

6.3 실험 2 42

6.3.1 데이터 44

6.3.2 비교 조건 44

6.3.3 실험 방법 45

6.3.4 결과 45

6.4 실험 3 46

6.4.1 데이터 46
6.4.2 비교 조건 46
6.4.3 실험 방법 47
6.4.4 결과 47

7 간호 진술문의 온라인 상태 값 측정법 49

7.1 실험 목적 49
7.2 데이터 50
7.3 실험 방법 51
7.3.1 상태 51
7.3.2 보상 53
7.3.3 절차 53
7.3.4 비교 조건 53
7.4 실험 결과 56

III 논의 및 결론 59

8 단어 감정 극성값 측정법 논의 61

8.1 TD 방법들 간의 분산의 차이 61
8.2 TD 기반 측정치의 평탄성 62
8.3 TD 방법의 점진성 63

9 간호 진술문 상태 값 측정법 논의 65

9.1 문장에 대한 함수 근사법의 문제 65
9.2 진술문 상태 값의 점진적 측정 68
9.3 진술문 상태들의 관계에 대한 TD 방법의 유연성 68
9.4 한계점 및 도메인 특정적 장점들 70

10 전체 논의 및 결론 73

10.1 희소한 보상 값의 문제 73

10.2 모형 기반 강화학습과 자연어 처리의 문제 74
10.3 본 연구에서 밝힌 것 76
10.4 강화학습을 이용한 자연어 처리의 미래 76

참고문헌 **79**
References 79

Abstract **87**

그림 차례

3.1 강화 학습에서 보는 에이전트와 환경과의 상호작용 모형 13

5.1 모형 기반 강화학습 주요 요소들의 관계도 27

5.2 차단 벽 미로(blocking maze) 과제에서 Dyna 모형들의 성능 그래프 28

5.3 TD-Gammon에 이용된 신경망 모형 30

6.1 6개 상태로 이루어진 마르코프 보상 과정의 예제 36

6.2 6개 상태로 이루어진 마르코프 보상 과정의 예제 2 36

6.3 TD 방법의 자질 선별 결과에 따른 감정 분류 정확도. 43

6.4 흔적 감쇄율에 따른 TD 방법(replacing traces)의 성능 그래프 . . 48

7.1 TextCNN 모형 55

7.2 학습 데이터에 대한 상태 값의 추이 그래프 58

8.1 Stanford Sentiment Treebank 데이터의 전체 어휘 셋에 대해 주석된 극성 값(annotated), 베이즈 확률 값(Bayesian), 그리고 TD(1) 기반 극성 값(significance traces)의 데이터 플롯 63

8.2 4개의 스필버그 영화에 대한 400개 영화평으로부터 단어 'spielberg'의 감정 극성 값 추적 그래프 64

9.1 Distributed Memory of Paragraph Vector (DM-PV 모형) 66

9.2 진술문의 시간적 연쇄에 따른 상태 값 그래프 70

표 차례

3.1 상태 전이 통제 여부와 완전한 관찰 여부에 따른 마르코프 모형 구분표 15

6.1 TD 방법의 최적 학습률과 흔적 감쇄율 42
6.2 TD 방법과 그 외 자질 선택 방법들에 따른 감정 분류 결과의 정확도 42
6.3 주석된 단어 극성값과 알고리즘을 통한 측정값과의 상관관계 . . . 45
6.4 LSTM 기반 보상 모형에 의존한 TD 방법과 IG 방법의 성능표 . . 48

7.1 간호 진술문 데이터 예제 50
7.2 약물 유해반응과 관련 있는 문장 주석의 예 51
7.3 간호 진술문의 상태 공간 지시 기호와 그 의미 52
7.4 SVM 학습에 사용된 최적 모수 값 54
7.5 TextCNN 학습에 사용된 모수 값 55
7.6 LSTM 분류기 학습에 사용된 모수 값 56
7.7 TD 기반 간호 진술문의 상태 값 57
7.8 ADR 분류 과제 정확도 실험 결과 58

8.1 학습률(0.1∼0.5)에 대한 TD(λ) 성능의 분산 평균값 61

9.1 DM-PV 모형을 이용한 문서 검색 수행도 67
9.2 간호 진술문 문장과 그 상태 기호 테이블 69
9.3 보상 감쇄율에 따른 상태 값 변화 테이블 70

10.1 강화학습 에이전트를 위한 Naive Bayes 모형의 작동 예제 75

10.2 감정 분류 과제 정확도 결과 75

제 1 장

서론

인간의 언어 습득 및 처리 과정은 인류의 오랜 수수께끼이다. 심리학에서 초기 강화 학습의 대표적인 연구자인 Skinner는 언어 학습을 자극(stimulus)과 강화(reinforcement)에 의해 설명할 수 있다고 보았고, 그의 유명한 저서 Verbal Behavior (Skinner, 1957) 에서 이를 주장하였다. 그리고 이 주장은 현대 언어학에 가장 중요한 영향을 끼친 언어학자인 Chomsky에 의해 곧 강하게 반박되었다. Chomsky (1959)는 강화만으론 인간의 언어학습을 설명할 수 없다고 지적했는데, 그 주요 이유로 아이들이 언어를 학습할만큼 충분한 자극에 노출되지 않고 (poverty of stimulus) 언어 사용에 대한 특별한 보상(reward)이나 처벌(penalty)이 없음을 들었다. Chomsky는 언어가 생득적(innate)이며 인간은 언어를 학습하는 능력을 타고 난다고 보았는데, 이는 강화 학습의 설명이 언어를 학습하기에는 지나치게 단순한 것임을 지적하는 맥락과 닿아 있다.

Chomsky의 해법은 그러나 언어학습을 위한 인지 모듈의 설명에 많은 부담을 안긴다. 그의 초기 통사론은 통사 규칙에 의존하여 통사처리 과정을 설명하는데 초점이 맞추어져 있었으나 그 기저의 구체적, 심리학적 처리모형에는 크게 관심을 두지 않았다. 그리고 그는 언어 모형에 대한 통계적 설명에 기본적으로 부정적이었다 (Chomsky, 1969). 하지만 인간의 언어 행동은 복잡하며 흔히 유사 규칙적(quasi-regular)인 패턴을 보이기 때문에, 오늘날 전산 언어학 혹은 자연어 처리 연구자들은 통계적 모형을 통해 언어를 모형화하는 것에 대해 거부감이 없다. 그러나 그것이 곧 우리가 그 설명에 충분히 만족한다는 것을 뜻하지는 않는다. 우

리는 계속해서 인간의 언어 지식을 인과적이며 인지적인 모형을 통해 성공적으로 설명하는 방법을 찾고자 노력한다.

강화 학습은 환경과 행위자 간의 상호관계 속에서 행위자에게 환경이 주는 보상을 매개로 하여 학습의 발생을 설명하고자 한다. 그리고 이는 동물의 원초적인 학습행위를 인과적이고도 컴퓨터 프로그램을 통해 재연 가능하게 설명하는데 있어 편리한 도구가 된다. 하지만 인간의 복잡한 언어 인지구조가 유전자적 측면에서 이미 주어져 있다면, 이에 대한 이론적 모형 없이 언어 학습 및 처리 양상을 설명하는 것은 이미 단순한 문제가 아니게 된다. 따라서 이 상황에서 우리가 취할 수 있는 하나의 전략은 인간의 언어 인지구조와 유사한 학습 환경을 가정하여 구축하고, 이를 배경으로 한 언어 처리 기제를 연구하는 것이다.

1.1 강화 학습의 짧은 역사

강화 학습(reinforcement learning)의 주요 개념은 행동 심리학 연구에서 동물들이 보상(reward)을 최대화시키기 위해 환경에 적응하는데 있어 자신의 경험을 어떻게 이용하는지를 실험적으로 연구하는 과정에서 나타났다 (Skinner, 1938; Thorndike, 1911). Thorndike는 그 유명한 고양이 상자 실험에서 고양이가 퍼즐 상자에서 빠져나오는데 걸리는 시간을 측정하였다. 고양이는 빠져나오는데 다양한 방법들 중 하나를 선택할 수 있었는데, 시행 수가 늘어날 수록 가장 빠르게 나오는 방법을 학습하고 선택하는 분명한 경향성을 보여주었다. 그리고 그의 선구적 연구를 통해 강화 학습의 중요한 두 가지 개념, 방법의 탐색(exploration)과 활용(exploitation)의 딜레마 해결이 중요하다는 것이 드러났다. 동물 행위자는 목표를 완수하기 위해 탐색을 통해 더 나은 해법을 찾는 동시에 찾은 해법을 문제 해결에 활용해야 한다. 학습과정에서 탐색을 너무 일찍 포기하면 더 나은 해결책이 존재함에도 이를 이용하지 못하게 되지만 탐색에 너무 많은 시간을 투자하는 것도 결과적으로 현명하지 않은 전략이다. 따라서 탐색과 활용 사이의 적절한 조화(balancing)가 강화 학습에서 중요한 문제가 된다. 또한 보상이 주어졌을 때, 이 보상을 얻기까지 취한 행동들에게 적절히 배분하는 문제(credit assignment problem)가 중요하다는 것이 알려졌다.

행동 심리학 안에서 강화 학습은 이후 여러 이론적 발전을 이룩했다. Skinner (1938)는 조작적 조건화(operant conditioning), 혹은 도구적 조건화(instrumental conditioning) 개념을 통해 학술적 개념들을 보다 명료화했다. 그에 따르면 행위자가 자신의 행동을 통해 환경에 변화를 가하고 그에 따른 보상에 따라 이로운 행동들을 학습하는 것은 조작적 학습에 속하는데, 왜냐하면 이 행동은 환경에 조작을 가하는 것이기 때문이다.

또 다른 중요한 이론적 발전은 Tolman (1948)에서 나왔다. 그는 강화 학습이 단순히 자극과 강화로 이루어진다는 관점이 실제와 맞지 않음을 실험을 통해 보여주었다. 그의 미로 실험에서 실험 쥐들은 초기에는 막혀서 학습하지 못했던 경로의 길이 다시 열렸을 때, 이를 재빨리 활용해서 전체 이동 경로를 바꾸는 행동을 취했는데 이는 단순한 자극과 강화 연쇄(S-R association)만으론 설명하기 어려운 것이었다. Tolman (1948)은 이를 설명하는 대안으로 "인지 맵(cognitive map)" 모형을 제안하였는데, 이 모형은 행위자가 주변의 물리적 특징에 대한 정보를 심적 표상(mental representation)으로 만들어 저장하고 이를 자신의 목표 성취를 위해 능동적으로 활용한다고 설명한다. 이와 같이 어떤 구조 혹은 모형을 강화 학습 모형 안에 도입하려는 시도는 기계학습 내에서는 흔히 계획 기반 강화 학습(planning-based reinforcement learning) 혹은 모형 기반 강화 학습(model-based reinforcement learning)으로 불린다.

1980년대 후반부터 강화 학습의 이론적 문제들을 배경으로 인공지능과 기계학습 연구자들은 강화 학습 이론들을 본격적으로 정초하기 시작하였다. 이 과정에서 기본 바탕이 된 중요한 이론적 기반은 Bellman (1957)의 마르코프 결정 과정(Markov Decision Process)으로 그는 이 모형을 동역학 계(dynamical system)에서의 에이전트의 행동을 최적으로 조정하는 문제(optimal control problem)에 적용하고자 하였다. 이러한 다이내믹 프로그래밍 방법론의 전통과는 별도로 강화 학습 특유의 시행착오(trial and error)를 통한 학습 모형은 Sutton and Barto (1981)에서 강화 학습 알고리즘의 특징으로 명시적으로 언급되었는데, 그들은 이 특징이 기계학습의 주요한 흐름인 지도학습(supervised learning)과 분명하게 다른 점으로 인식하였다. 그리고 80년대에 등장한 시간차 학습(temporal-difference

learning)의 출현(Sutton, 1984, 1988) 또한 강화 학습의 역사에서 중요한 이정표 중 하나로, 시간차 학습 알고리즘은 Watkins의 Q-learning (Watkins & Dayan, 1992), Sarsa (Sutton, 1996), REINFORCE (Williams, 1992) 그리고 Actor-critic 모형 (Witten, 1977; Barto, Sutton, & Anderson, 1983)과 밀접한 관계를 가지며 강화 학습 알고리즘의 스펙트럼을 크게 넓히는 역할을 하였다.

비록 이론적 발전은 이미 2000년 이전에 많이 이루어졌으나 21세기에 들어와 컴퓨터의 하드웨어적 발전과 인터넷의 등장으로 인한 데이터의 범람은 많은 기계학습 방법론의 폭넓은 응용을 가능하게 하였는데, 특히 강화 학습 알고리즘이 실험실에 한정된 문제를 넘어서 인간의 인지능력과 맞먹는 성능을 보여준다는 것이 널리 알려졌다. Mnih et al. (2015)는 심층 신경망을 적용한 Deep Q-network 모형을 이용한 학습 에이전트를 Atari 사의 전통적인 전자오락 게임들에 도전시켜 인간 전문가 플레이어의 수행을 뛰어넘는 성능을 기록하였고, Silver et al. (2016, 2017)은 Monte-carlo tree search 알고리즘을 응용한 강화 학습 알고리즘으로 바둑경기를 학습시켜 공식 바둑 경기에서 인간 최고수를 압도하는 성능을 성공적으로 시연하였다.

1.2 연구 질문들

심리학에서 출발한 강화 학습은 기계학습 분야에서 하나의 학습 패러다임으로 자리 잡으며 크게 발전하고 있다. 그러나 강화 학습 방법론은 robotics와 같은 전통적인 optimal control 문제에서 벗어난 문제에는 여전히 적용이 생소한 학습 방법론으로 남아 있다. 또한 자연어 처리 혹은 언어 학습과 같은 복잡한 문제영역에서 강화 학습을 이용한 접근법이 얼마나 효과적일 수 있을지도 현재는 분명하지 않다.

강화 학습을 어떻게 자연어 처리 연구에 응용할 수 있는가? 바로 이 질문에 대해 응답하려는 욕구가 이 논문의 가장 큰 기본적인 동기이자 배경이다. 하지만 자연어 처리는 매우 방대한 분야로 많은 작은 주제들과 그에 대한 해결책들의 집합으로 이루어져 있으므로, 본 연구는 비록 작지만 구체적이고, 또한 다른 문제에도 확장 가능한 아래의 두 가지 확장적 주제에 초점을 맞춘다.

- 텍스트의 부분 단위의 연속적 상태에 대한 가치 예측의 문제
- 텍스트의 연속적 상태에 대한 가치 예측 및 유형 분류의 문제

전자의 과제는 텍스트 단어들의 연속적 감정가(sentiment polarity value) 측정이라는 문제로 다루어졌고, 후자의 과제는 어떤 대상의 상태를 기술하는 텍스트가 반영하는 상태의 가치 및 그 상태의 유형에 대한 판단의 문제로서 본 논문에서 구체화되었다. 본 논문의 실험은 병원에 입원한 환자들을 그 대상으로 삼았으며 환자에 대한 간호 진술문들을 환자의 상태를 반영하는 텍스트로 정의하였다. 따라서 이 텍스트, 즉 간호진술문은 환자라는 대상의 특정 시점의 상태에 대한 가치를 반영하기에 이 상태 값을 예측하고 그 상태 범주를 판단하는 것은 환자라는 대상에 대한 유의미한 정보를 얻는데 유용하다 할 수 있다.

위 두 과제를 강화 학습 알고리즘을 통해 접근할 때, 우리가 얻게 되는 중요한 이점은 **연속적 분석 가능성**에 있다. 일반적인 지도 학습법에서와 달리 강화 학습 알고리즘은 학습시기(training phase)와 시험시기(test phase)가 따로 있지 않고, 연속적이며 점증적인(incremental) 처리를 지향한다. 따라서 기본적으로 강화 학습 알고리즘은 "online" 처리 모형이므로, 강화 학습을 이용한 위 접근법은 우리가 관심 있는 대상에 대한 연속적 상태 변화 관찰에 적절하다.

1.3 논문의 전체적 구성

본 논문은 소개말 섹션 이후에 먼저 연구의 전체적 배경을 설명한다. 언어 처리의 점진적 성격에 대한 언어 심리학적인 연구를 요약하며 인간의 언어이해가 본질적으로 연속적이며 점진적 학습을 통해 이루어진다는 점을 소개할 것이다. 그리고 강화 학습의 주요 개념을 기술한 후에, 가치 예측의 문제에 널리 적용되는 강화 학습 알고리즘인 시간차 학습(temporal-difference learning)을 설명한다. 시간차 학습 방법은 본 연구에서 단어, 문장의 가치 예측의 문제에 모두 중요하게 이용되었기에, 그 중요성을 감안하여 다른 방법에 비해 상대적으로 자세하게 서술되었다.

파트 II는 단어 감정 극성값 예측 문제에 시간차 학습 알고리즘인 TD(λ)를 적용한 실험과 간호 진술문 텍스트의 온라인 상태 값 예측 및 분류 과제 실험으로

이루어져 있다. 그리고 마지막 파트 III은 실험의 결과와 그 의미를 논의한다.

이미 강화 학습 내용에 익숙한 독자라면 파트 II의 강화 학습 관련 내용들을 건너 뛰어도 이해에 무방하다. 그리고 만일 독자의 관심이 강화 학습을 이용한 본 연구의 이론적 측면에만 관심이 있다면 마지막 파트인 논의 및 결론 부분을 볼 것을 권고한다.

제 I 편

연구 배경

제 2 장

언어 처리의 점진성

문장이 처리되고 그 의미가 이해되는 과정에 대해서 중요한 의문이 하나 있다. “문장을 전부 읽은 다음에야 비로소 문장의 의미가 처리되는 것인가?” 아니면 “문장을 읽어 가면서 점차적으로 이해가 일어나는가?” 이 질문에 대한 답을 생각하기 전에 인식해야 할 한 가지 분명한 사실은 우리가 문장의 의미를 제대로 이해하기 위해서는 그 문장의 통사적 구조를 파악해야 한다는 것이다. 그리고 그 이후에야 그 문장의 의미를 제대로 이해할 수 있다. 통사론적인 분석을 해보면 문장이 단순한 단어의 연결이 아니라는 사실은 쉽게 이해된다. 어떤 단어들은 서로 거리가 멀리 떨어져 있어도, 영어에서 주어의 단수/복수가 동사의 형태에 미치는 경우에서와 같이 서로 밀접한 문법적 관계를 갖는다. 즉, 문장은 통사적 나무 구조를 갖고 있고, 이에 따라 그 의미가 이해될 수 있다. 문제는 그 구조가 문장을 읽을 때, 언제 어떻게 형성되는가이다.

2.1 통사적 처리의 점진성

만일 문장의 통사처리가 점진적으로 이루어지지 않고 문장 읽기를 끝마치는 시점에서 한 번에 이루어진다고 가정해보자. 그렇다면 작업 기억(working memory)은 문장의 구조 분석에 필요한 모든 정보를 문장 읽기를 마칠 때까지 저장하며 유지해야 할 것이다. 그러나 이 전략은 작업 기억에 큰 부담을 지우는 전략임을 쉽게 알 수 있다.

언어 심리학의 연구결과들은 문장의 통사 분석이 어느 특정 시점에서 이루

어지는 것이 아니라 각 단어들을 인지하자마자 곧바로 분석에 활용한다는 "해석의 즉시성(immediacy of interpretation)" 전략을 지지한다 (Marslen-Wilson, 1973, 1975). 이 전략을 지지하는 많은 실험적 증거들은 일시적으로 통사 구조적으로 애매성(ambiguity)을 갖는 문장들, 이른바 "뜰길 걷기 문장(garden-path sentences)" 문장들에 대한 읽기 시간을 측정함으로써 나왔다.

(1) The defendant examined <u>by the lawyer</u> turned out to be unreliable.

문장 (1)의 'examined'는 일시적으로 중의적인 통사적 해석을 일으킬 수 있는데 왜냐하면 'by the lawyer'를 보기 전까지 읽는 이는 그 동사가 관계절의 동사가 아니라 주어의 동사라고 해석하기 쉽기 때문이다. 여러 읽기 실험 연구에서 사람들은 'by the lawyer'과 같은 관계절이 드러나는 부분에서 공통적으로 읽기 시간이 증가하는 패턴을 보였다 (Rayner, Carlson, & Frazier, 1983; Clifton Jr et al., 2003; Ferreira & Clifton Jr, 1986). 이 현상은 사람들이 초기에 구성한 통사 해석을 포기해야 했다는 것을 가리키며, 바꿔 말하면 사람들이 단어를 봄과 동시에 어떤 특정 해석을 선호하고 있었음을 함축하는 결과로 이해된다.

2.2 의미적 처리의 점진성

사람이 텍스트의 의미를 이해하기 위해서는 문장의 통사 구조 뿐 아니라 지시체와 사건들에 대한 심적 표상 혹은 정보, 그리고 선행사들의 지시 관계 등에 대한 것들을 인지적으로 파악해야 한다. 따라서 가장 기본적으로는 대명사가 앞서 나타난 어떤 대상을 지시하는지와 같은 정보를 알아야 한다.

안구운동 추적을 통한 여러 실험에서 대명사 지시 관계에 대한 사람들의 읽기 시간을 측정한 결과, 선행사가 성별 관련 문법 제약을 위반할 때(e.g. 'herself'가 'the pilot'을 지시하는 경우) 읽기 시간이 증가하는 현상을 관찰하였다 (Duffy & Keir, 2004; Sturt, 2003; Sturt & Lombardo, 2005). 또한 지시자와 지시체의 공간적 거리가 길어질수록 안구 고정 시간(fixation time)이 증가하는 것을 관찰하였다 (Ehrlich & Rayner, 1983; Garrod, Freudenthal, & Boyle, 1994). 그 외에도 선행사가 저빈도(low-frequency) 단어이거나 대명사가 연접된 명사구 (conjoined

NP)의 한 구성소를 가리키는 경우에도 안구 고정 시간이 증가한다는 것이 드러났다 (Van Gompel & Majid, 2004; Albrecht & Clifton, 1998).

(2) He stabbed her with his knife.

(3) He assaulted her with his weapon.

(4) He threw the knife into the bushes, took her money, and ran away.

사람들은 또한 추론 문제에 있어서도 즉각적인 온라인 처리를 선호한다는 것이 한정 명사구(definite NP)와 선행사의 관계 양상에 대한 실험에서 관찰되었다 (O'brien, Shank, Myers, & Rayner, 1988). 독자들은 문장 (4)를 보기 전에 문장 (2) 혹은 문장 (3) 만을 볼 수 있었는데, 문장 (3)을 먼저 본 실험 참가자들은 문장 (2)를 본 참가자들에 비해 문장 (4)의 목표어 'the knife' 처리에서 더 많은 읽기 시간을 기록했다. 이는 사람들이 지시어가 가리키는 대상의 추론을 미루지(delay) 않고 즉각적으로 수행한다는 것을 의미하는 것이다. 그 외 다른 후속 연구에서도 위의 목표어의 유형이나 의미역이 선행하는 맥락과 일치할 경우, 해당 목표어의 처리시간이 더 빨라지는 현상을 관찰함으로써 텍스트 독해시 의미 처리의 점진성을 실험적으로 뒷받침하였다 (Myers, Cook, Kambe, Mason, & O'Brien, 2000; Cook & Myers, 2004; Garrod & Terras, 2000).

제 3 장

강화 학습

본 장을 포함한 3~5장의 주요 내용은 강화 학습의 일반적인 내용을 요약하며 제공하는 것이다.[1] 기계 학습 안에서 강화 학습은 지도 학습(supervised learning)과 비지도 학습(unsupervised learning)과 구별되는 독특한 위치를 가지고 있다. 그리고 그 근본적인 배경에는 강화 학습 특유의 이론적 프레임이 있다. 강화 학습이 학습문제를 바라보는 관점은 환경 안의 에이전트가 처벌(punishment)을 피하면서 보상(reward)을 최대화 할 수 있도록 의사 결정(decision making)을 하는 법을 배울 수 있는지에 대한 이론적 모형으로 이루어진다. 이 프레임워크 아래에서 다양한 적응적 학습 알고리즘 및 기제들을 모사하는 알고리즘들이 고안 되었다.

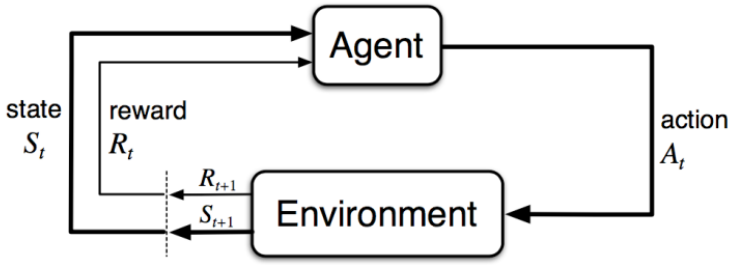


그림 3.1: 강화 학습에서 보는 에이전트와 환경과의 상호작용 모형

3.1 강화 학습 모형

그림 3.1 (Sutton & Barto, 1998, p. 52) 에서 볼 수 있듯이, 강화 학습에서 행위 주관자인 에이전트는 환경과 상호작용을 하며 학습한다고 가정된다. 여기서 일어나는 상호작용은 시간 선상에서 일어나는데, 먼저 에이전트가 시간 t 시점에서 행동 A_t를 행하면 환경(environment)은 그에 대한 반응으로 상태 S_{t+1}로 바뀌고 그에 따른 보상 R_{t+1}을 에이전트에게 보낸다. 그러면 에이전트는 그 신호를 환경에서 주는 피드백으로 보고 그에 따라 자신의 다음의 행동을 결정하게 되고 이 과정은 계속 반복된다. 에이전트는 이 과정에서 스스로 보상을 최대화 할 수 있는 행동 체계를 학습할 수 있는데, 여기서 일어나는 학습이란 변화하는 상태에 대한 에이전트의 가치 함수와 자신의 행위에 대한 가치 함수의 수렴을 의미한다. 이 모형에서 환경의 변화는 결정적이지 않고 확률적(stochastic)이라고 흔히 가정된다.

3.2 마르코프 결정 과정

강화 학습 모형의 특징들은 마르코프 결정 과정(Markov Decision Process: MDP) 모형을 통해 잘 기술된다. 마르코프란 이름에서 쉽게 유추할 수 있듯이 마르코프 결정 과정은 미래 상태를 예측하는데 있어 현재 상태의 정보만으로 충분하다는 "마르코프 속성 원리"를 받아 들인다. 마르코프 결정 과정으로 모델링하고자 하는 환경의 상태 수가 유한할 경우, 마르코프 상태는 상태 이행 확률표로 알기 쉽게 표현할 수 있다 (수식 3.1).

$$P_{ss'} = P[S_{t+1} = s' | S_t = s] \tag{3.1}$$

그리고 마르코프 결정 과정(MDP)은 형식적으로 총 5개의 구성 요소를 가진 튜플, (S, A, P, R, γ)로 정의되며 각 요소는 아래와 같은 의미를 가진다.

- S는 상태들로 이루어진 유한 집합이다.
- A는 유한한 수의 행동들의 집합이다.

[1]따로 전거를 언급하지 않는 이상 강화 학습 알고리즘에 대한 내용은 기본적으로 Sutton and Barto (1998)와 Szepesvári (2010)에 기반한다.

Markov models		상태 이행의 통제 가능 여부	
		No	Yes
상태의 완전한 관찰 가능 여부	Yes	Markov Chain, MRP	MDP
	No	Hidden Markov Model	POMDP

표 3.1: 상태 전이 통제 여부와 완전한 관찰 여부에 따른 마르코프 모형 구분표

- P는 상태 이행 확률표이다.
- R은 상태 S_t에 따라 실수 보상 값, $r \in \mathbb{R}$을 리턴하는 보상 함수이다.
- $\gamma \in [0, 1]$는 보상에 대한 감쇄 비(discount factor)이다.

MDP에서 감쇄 가중치(discount factor)를 쓰는 사용하는 이유는 크게 두 가지이다. 첫째 이유는 수학적인 고려에 의한 것으로 γ가 1 미만일 때, 그것의 무한 급수는 기하 급수(geometric series)가 되고 기하 급수는 수렴 급수이기 때문에 ($\sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$) 이 가중치에 의해 곱해진 결과 값의 합이 발산하지 않도록 돕는다. 둘째 이유는 심리적/인지적 요인에 의한 것으로, 동물들은 보다 가까운 미래의 보상을 먼 미래의 보상에 비해 크게 평가한다는 심리학적 직관에 근거하고 있다 (Myerson & Green, 1995; Frederick, Loewenstein, & O'donoghue, 2002).

마르코프 결정 과정이 에이전트와 환경 과의 상호작용에 대한 유일한 모형은 아니며 주어진 강화 학습 문제에 따라 다양한 형태로 변형 혹은 확장이 가능하다. 일반적인 MDP에선 환경이 주는 정보를 에이전트가 완전하게 인식 가능하다고 가정하지만, 그렇지 않은 경우도 모델링할 필요가 생길 수 있다. 이런 경우를 모형에 반영한 MDP를 partially observable MDP(POMDP)라고 부르며, 이 모형에서 에이전트는 환경이 주는 정보를 항상 정확하게 인지하지 못한다고 가정된다. 이것을 정리하여 상태의 완전한 관찰 가능성과 상태 전이과정에 초점을 맞추어 MDP를 크게 분류하면 표 3.1와 같다. 비록 본 논문의 범위를 벗어나는 주제이긴 하나 환경의 상태가 이산적이지 않고 연속적일 경우(continuous MDP), 상태가 오직 하나인 경우 등 다양한 경우에 대응하는 MDP 모형들이 존재한다 (Sutton & Barto, 1998).

3.3 목적, 상태, 행위 그리고 보상

상술하였듯이 강화 학습 환경에서 에이전트의 목적은 단기적 측면 뿐 아니라 장기적 측면에서의 보상을 최대화하는 것이다. 형식적으로 표현하면, 각 시간 t에서 보상 r_{t+1}를 얻을 때 에이전트의 학습 목적은 미래의 보상 기대 값을 최대화하는 것이다 (수식 3.2).

$$\mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}\right] \tag{3.2}$$

MDP에서 상태(state)란 매우 추상적이면서도 일반적인 용어로 쓰이는데 이는 에이전트가 처할 수 있는 매우 다양하지만 구별 가능한(distinct) 상황의 집합을 가리킨다. 때문에 상태는 공간적인 위치를 표상할 때는 GPS 그리드 같은 실수 값의 순서 쌍으로 나타날 수도 있고, 어떤 상태의 긍정 혹은 부정 상태를 가리키고자 할 때는 0 혹은 1 처럼 이진 값으로 표현될 수도 있다. 또한 상태는 실수 값으로 표현되는 경우처럼 연속적 상태일 수도 있고, 이진 값으로 표현될 때처럼 이산적 상태일 수도 있다. 그리고 만일 상태가 연속적이라면, 테이블과 같은 자료 구조로는 이를 표상할 수 없으므로 연속적 상태를 보다 적은 차원 수로 근사적으로 표상해주는 함수에 의존하는 전략을 사용하게 된다. 이를 강화 학습 진영에서는 흔히 함수 근사 접근법(function approximation approach)이라고 부른다.

행동은 에이전트가 각 상태 혹은 상황에서 결정 가능한 선택지들을 가리킨다. 물론 이 선택 가능한 행동들의 집합이 항상 모든 상태에서 동일한 것은 아니다. 예를 들어, 정사각형 칸들로 이루어진 2차원 미로와 같은 환경을 가정해보자. 그리고 이 환경에서 에이전트는 각 칸에서 동, 서, 남, 북 중 어느 한 방향으로 한 칸 씩 이동할 수 있다고 하자. 그렇다면 에이전트는 미로의 가장자리 혹은 한 쪽면 끝에서는 네 가지 가능한 이동 선택지를 가질 수 없기 마련이다. 이와 같이 선택 가능한 행동은 에이전트가 처한 상태에 따라 제한될 수 있다. 분명 이런 경우는 환경에 의한 행동의 제약이라고 볼 수 있으나, 생각해 보면 반드시 이 제약이 에이전트에 대해 외부적일 필요는 없다. 즉, 에이전트 내부적 요인에 의한 행동 제약도 충분히 가능하다. 예를 들어, 비밀번호를 잊어버려 들어가야 할 방의 출입문을 열지 못하는 경우처럼, 사람은 기억 능력의 제약으로 인한 행동의 제약을

흔히 경험한다. 또한 상태에 대한 표상이 연속적일 수 있듯이, 행동 역시 셀 수 있는 선택지가 아닌 무한히 많은 경우를 생각해야 할 수 있다. 그리고 이 경우에도 상태에서와 마찬가지로 함수 근사 방법을 통해 행위 공간을 근사하여 표상하는 접근을 취하게 된다.

보상(reward)은 형식적으로 시간 t에서 에이전트에게 주어지는 스칼라 실수 값으로, 에이전트가 그 시점에서 얼마나 잘 지내고 있는지를(well-being) 반영하는 역할을 한다. 상태 개념과 마찬가지로 보상 개념은 다양한 형태로 구체화될 수 있다. 볼펜을 손가락 끝에서 세우는 과제 같은 경우, 볼펜이 균형을 잃어 떨어지지 않을 때마다 양수 값을 보상으로 주고 떨어질 경우는 음수 값을 줄 수 도 있다. 미로 탈출 과제 같은 경우에는 오직 출구를 찾아 나갈 경우에만 양수 값을 보상으로 받고, 그렇지 않고 미로 안에서 머무는 모든 시간마다 일정한 양의 음수 값을 받게 할 수도 있다. MDP에서는 또한 보상이 반드시 결정적(deterministic)이 아니라 확률적으로 주어질 수도 있다. 그리고 강화 학습 프레임워크에서 가정 하듯이 에이전트의 목적은 바로 이 보상의 기대 값을 최대화하는 것이 된다.

보상이라는 개념은 이해하기는 쉬우나 실제 문제에서 모델링 하기에는 쉽지 않을 수 있다. 왜냐하면 에이전트의 학습 목적 자체가 보상 값에 의존하기 때문에, 값의 범위를 적절하게 결정하지 않은 경우, 에이전트의 예측하기 힘든 행동을 유발할 수도 있기 때문이다. 한 가지 쉬운 예로 자동차 경주와 같은 과제에서 경주 트랙을 벗어날 경우에 -100과 같은 매우 큰 처벌 값을 받는 반면, 트랙 안에 머무는 동안에는 -1 점을 받는다고 해보자. 이런 학습 환경의 기획자는 처음에 이런 보상 시스템이 자동차가 절대 트랙을 벗어나지 않게 안전한 운행을 하게 할 것이라고 생각할지도 모른다. 하지만 그 예상과는 달리 자동차는 트랙에 빠지게 될 경우에 얻는 손실이 너무 크기 때문에 앞으로 나가며 길을 찾는 시도를 포기하고 제 자리에 머무는 선택지를 택할 수도 있다.

3.4 강화 학습 알고리즘의 두 특징

MDP 문제 해결을 위한 초창기 주된 해결책은 다이내믹 프로그래밍이었으나, 이 알고리즘은 매우 적은 수의 상태와 행동들을 가지는 단순한 문제에만 적용

가능하다는 문제점이 있었다 (Bellman, 1957; Sutton & Barto, 1998; Szepesvári, 2010). 강화 학습 알고리즘은 다이내믹 프로그래밍의 이 치명적 단점을 극복할 수 있도록, 아래와 같은 두 가지 전략을 취한다 (Szepesvári, 2010).

- 샘플링(sampling)을 활용하여 기대 값을 추정한다.
- 함수 근사 방법을 통해 가치 함수를 표상한다.

샘플링을 이용한 가치 추산은 복잡한 현상에 대한 효율적인 학습을 가능하게 해준다는 큰 장점이 있으며, 이를 통해 현상 기저의 구조에 대한 사전 지식을 불요하게 만든다는 무시 못할 편리함을 제공한다. 두 번째 특징은 모든 강화 학습 방법이 취하는 전략은 아니나, 강화 학습이 실제적인 문제에 대한 효율적인 대안이 되기에 매우 중요한 요소로 꼽힌다. 왜냐하면 우리가 실제로 어렵다고 느끼는 많은 문제들의 상태 공간 혹은 가능한 행위의 종류는 매우 많은 것이 보통이기 때문이다. 위 두 가지 전략은 복잡한 상황에서 에이전트가 부딪히는 문제의 경우의 수를 크게 줄여주는 역할을 하는데, 왜냐하면 샘플링을 통해 매우 큰 상태 공간의 일부분만을 학습의 대상으로 삼을 수 있으며, 함수 근사법을 통해 그 일부분의 차원을 더 나아가 크게 감소시킬 수 있기 때문이다.

제 4 장

가치 예측 문제에 대한 시간차 학습

본 장에서는 강화 학습의 핵심적인 알고리즘의 하나인 시간차 학습(temporal-difference learning)에 대해서 기술한다. 이를 위해 먼저 시간차 학습의 기본적인 환경이라고 할 수 있는 마르코프 보상 과정(Markov Reward Process: MRP)을 소개한 후에, 가치 함수의 정의를 설명한다. 강화 학습에서 가치 함수는 크게 상태 가치 함수(state value function)와 행위 가치 함수(action value function)로 구분되고, 이들의 수렴 가능성은 Bellman update operation에 의해서 수학적으로 보장된다. Sutton (1984, 1988)의 TD(λ) 알고리즘은 가치 예측 문제에 있어서 몬테 카를로 방법과 시간차 방법 모두를 적격 흔적도(eligibility traces) 개념을 통해 하나로 통합한 알고리즘으로 이 장의 마지막에 서술 되었다.

4.1 마르코프 보상 과정

마르코프 보상 과정(MRP)은 마르코프 결정 과정의 부분 집합이라고 볼 수 있으며 총 4개의 구성 요소를 가진 튜플, (S, P, R, γ)로 정의된다. 그리고 그 구성요소의 의미는 아래에서 볼 수 있듯이 마르코프 결정 과정과 동일하다. 마르코프 보상 과정은 기본적으로 마르코프 연쇄(Markov chain) 모형과 매우 유사하며, 따라서 이 환경에서 에이전트는 상태 이행에 대한 조절 기제를 가지지 않는다.

- S는 상태들로 이루어진 유한 집합이다.
- P는 상태 이행 확률표이다.

$P_{ss'} = P[S_{t+1} = s'|S_t = s]$

- R은 상태 S_t에 따라 실수 보상 값, $r \in \mathbb{R}$을 리턴하는 보상 함수이다.
- $\gamma \in [0, 1]$는 보상에 대한 감쇄 비(discount factor)이다.

4.2 가치 함수

가치 함수는 미래 보상에 대한 예측을 제공하는 기능을 하므로 상태들에 대한 가치 평가를 하는데 유용하다. 여기서 강화 학습의 목표는 최적의(optimal) 가치 함수를 얻는데 있다고 할 수 있는데, 가치 함수가 최적 함수라는 것을 표기시, $V^*(s), s \in S$로 표시한다. 이 함수의 의미는 어떤 상태 s에서 에이전트가 얻을 수 있는 정확한 기대 값을 뜻한다.

MRP에서는 다음 상태 s_{t+1}와 보상 값 r_{t+1}은 상태 전이 확률 함수, $P(\cdot|S_t)$에 의해 결정된다. 그리고 MRP는 시간 T의 길이를 갖는 어떤 일화(episode)에 대한 감쇄된 보상 값(discounted return)을 아래의 수식을 통해 정의한다.

$$G_t = \sum_{k=0}^{T-t} \gamma^{k-1} R_{t+1} \tag{4.1}$$

그리고 MRP는 수식 4.2를 통해 각 상태 $s \in S$에 대한 보상의 기대 값을 제공하는 가치 함수 V를 학습시키는데 이용된다.

$$V(s) = \mathbb{E}[G_t|S_t = s] \tag{4.2}$$

4.3 몬테 카를로 방법

몬테 카를로 방법(Monte-Carlo methods)은 샘플들 값의 평균을 기반으로 MDP를 해결하려는 접근법을 가리킨다. 이 방법에서 에이전트는 한 일화(episode)가 완료되어야만 그로부터 가치 함수를 업데이트할 수 있으며, 일화가 종료되기 전에는 상태나 행위에 대한 값을 얻을 수는 있는 길이 없다. 그러므로 몬테 카를로 방법은 일화 단위로서만 점진적(incremental)인 방법으로 여겨진다.

한 일화로부터 얻을 수 있는 감쇄된 보상 값(discounted return)은 수식 4.1과 마찬가지로 아래와 같이 정의 되므로,

$$G_t = R_{t+1} + R_{t+2} + \cdots + \gamma^{T-1} R_T \tag{4.3}$$

어떤 상태 s에 대한 가치 함수는 수식 4.2과 같다.

하지만 평균이 다음의 공식, $\mu_k = \mu_{k-1} + \frac{1}{k}(x_k - \mu_{k-1})$을 통해 정진적인 방식으로 계산될 수 있듯이 이 가치 함수 또한 수식 4.4를 통해 점진적으로 계산 가능하다.($cnt(S_t)$는 상태 S_t의 출현 횟수 카운터를 의미한다)

$$V(S_t) = V(S_t) + \frac{1}{cnt(S_t)}(G_t - V(S_t)) \tag{4.4}$$

그리고 우리가 이 가치 함수를 이미 경험한 일화들과 상관없이 지속적으로 학습하는 함수로 만들고자 한다면 카운터 함수($cnt(S_t)$)를 $0 < \alpha < 1$의 학습률 파라미터로 대체할 수 있다. 이것을 표현한 것이 수식 4.5이며, 몬테 카를로 방법에 의한 가치 함수를 가리킨다.

$$V(S_{t+1}) = V(S_t) + \alpha(G_t - V(S_t)) \tag{4.5}$$

4.4 시간차 예측

시간 차 학습(temporal-difference learning: TD)은 강화 학습의 대표적인 알고리즘으로 다이내믹 알고리즘과 몬테 카를로 방법을 통합하는 방법론으로 평가 받는다 (Sutton, 1984, 1988). 시간 차 학습의 목표는 경험 내용에 따라 시점 t에서의 상태 S_t에 대한 상태 값을 측정하는 것이다. 시간 차 학습 방법은 본질적으로 점진적(incremental)이라는 특징을 갖는데, 이는 상태 값들을 갱신하는데 있어 종결된 일화(episode)를 요구하지 않고 모든 시간 지점에서의 갱신이 가능하기 때문이다. 또한 시간 차 방법은 가치 예측에 있어서 부트스트래핑(bootstrapping)을 이용한다는 특징을 갖는데, 이는 시간 차 방법이 다이내믹 프로그래밍처럼 마르코프 결정 모형을 이용하여 상태 값을 측정하기 때문이다.

Sutton (1981, 1984)의 TD(λ) 알고리즘은 적격 흔적도(eligibility trace) 개념을 도입하여 흔적 감쇄율(trace-decay rate), λ 파라미터를 통해 가장 단순한 형태의 TD 방법과 몬테 카를로 방법을 점진적 학습(incremental learning)이라는 접근 패러다임 안에서 통합한 것으로 평가 받는다. 파리미터 λ는 0에서 1까지의 실수 값을 갖는데, 가장 단순한 시간차 방법인 TD(0)의 가치 함수 갱신 규칙은 아래와 같다. (α는 학습률을 가리킨다)

$$V(S_t) = V(S_t) + \alpha(R_{t+1} + \gamma V(S_{t+1}) - V(S_t) \tag{4.6}$$

만일 λ가 1이라면 이는 '매 방문 몬테 카를로 방법(every-visit MC)'과 동일해지는데, 이것은 곧 어떤 상태가 일화 안에서 출현 할 때마다 평균 리턴 값을 계산하기 위한 방문 카운터와 리턴 값과 함께 증가한다는 것을 의미한다. 따라서 TD(λ) 모형은 흔적 감쇄율이라는 파라미터를 조절함으로써 TD와 MC 사이를 매우 자연스럽게 왕복할 수 있다. 일반적으로 알려진 TD(λ) 알고리즘은 알고리즘 1과 같다 (Sutton & Barto, 1998).

Algorithm 1: TD(λ) with replacing traces

```
Initialize V(s) arbitrarily and let e(s) = 0 for all s ∈ S;
repeat
    while s_t not at end of the episode do
        observe reward, r, and s_{t+1};
        δ ← r + γV(s_{t+1}) − V(s_t);
        e(s_t) ← 1;
        for all s ∈ S do
            V(s) ← V(s) + αδe(s);
            e(s) ← γλe(s);
until the episode is terminal;
```

TD(λ) 알고리즘은 시스템의 상태 수가 많을 경우 상당히 큰 계산속도 상의 비용을 요구한다. 하지만 다행히 적격 흔적도 벡터는 일반적으로 대부분 0으로 채워져 있기 때문에 희박한(sparse) 벡터이며, 따라서 본 연구에서는 이를 수정한 알고리즘 2를 사용하였는데, 만일 주어진 데이터의 토큰 수가 N이라면 이 수정된

알고리즘의 상태 값 계산 복잡도는 $\mathcal{O}(N)$이 된다.[1]

Algorithm 2: Fast TD(λ) with replacing traces

```
Initialize V(s) arbitrarily and let e(s) = 0 for all s ∈ S;
H ← new hash table;
repeat
    while s_t not at end of the episode do
        observe reward, r, and s_{t+1};
        δ ← r + γV(s_{t+1}) − V(s_t);
        e(s_t) ← 1;
        if H not contains s_t then
            insert s_t into H;
        for all h ∈ H do
            if e(h) ≤ 0.001 then
                e(h) ← 0;
                remove h from H;
                continue;
            V(h) ← V(h) + αδe(h);
            e(h) ← γλe(h);
until the episode is terminal;
```

TD(λ) 모형에는 여러가지 유형의 적격 흔적도(eligibility trace)가 있고 가장 통상적으로 쓰이는 흔적도는 'replacing' 흔적도로 알고리즘 1과 2와 함께 제시되었다. 'accumulation' 흔적도는 그 이전 흔적도 값에 1을 더해 누적시키는 방법의 흔적도로 정의는 4.7과 같다.

$$e(S_t) \leftarrow e(S_t) + 1 \tag{4.7}$$

만일 $\lambda = 1$이고 'accumulation' 흔적도를 사용한다면, 이는 곧 'every-visit MC' 방법과 동일하다. 그리고 $\lambda = 1$이면서 흔적도 방법으로 'replacing'을 선택한다면, 이것은 일화 내에서 처음 등장 했을 때만 값을 갱신하는 'first-visit MC' 방법과 같아진다.

최근에는 True Online TD(λ) 모형을 통해 'Dutch' 흔적도라는 새로운 유형도 제시되었다 (Seijen & Sutton, 2014; Van Seijen, Mahmood, Pilarski, Machado,

[1]가정된 학습 환경은 마르코프 보상 과정이므로 행위 선택 부분이 없다.

& Sutton, 2016). 이 흔적도의 정의는 수식 4.8와 같다.

$$e(S_t) \leftarrow (1-\alpha)e(S_t) + 1 \tag{4.8}$$

어떤 적격 흔적도를 사용하고, 어떤 λ 값을 사용하는 지는 이론적인 문제의 영역이 아니라 경험적인 문제에 가깝다. 통상적으로는 'replacing' 흔적도가 널리 사용되고 성능이 더 나은 것으로 보고 된다 (Sutton & Barto, 1998). 또한 어떤 흔적 감쇄율(λ)을 사용할 것인지에 대해서는 문제 환경이나 방법에 따라 결정하기 복잡한 문제이나 보통 0 이상의 값이 선호된다.

제 5 장

모형 기반 강화 학습과 함수 근사법

강화 학습에서 모형(model)이란 행위자가 자신의 행위에 대한 환경의 반응을 예측하는데 활용가능한 모든 것을 가리킨다. 그리고 모형 기반(model-based) 강화 학습은 모형 자유(model-free) 강화 학습과 대비되는 말로 흔히 이해 되는데, 이 차이는 모형 기반 방법이 학습(learning) 보다는 계획(planning)에 초점을 둔다는 점에서 비롯된다. 하지만 이 둘은 상태에 대한 가치 함수에 깊이 의존한다는 점에서 큰 공통점을 가지기도 한다. 본 장에서는 모형 기반 강화 학습의 일반적 방법론과 더불어 근사 함수를 통한 상태 표상 방법, 그리고 마지막으로 인지 모델링을 이용한 가치 함수를 통해 인지모형 기반 강화 학습 환경 구축의 가능성을 언급하고자 한다.

5.1 모형과 계획

모형 기반 강화 학습에서 모형은 모형 자유 강화 학습에서의 '환경' (environment)을 모사(simulate)하는 역할을 수행한다. 만일 모형이 가능한 모든 상태들에 대한 확률 분포를 제공한다면 이것을 분포 모형(distribution models)이라고 부르고, 단지 어떤 분포 모형에 의한 결과물만을 제공한다면 샘플 모형(sample models)이라고 부른다. 그리고 마르코프 결정 과정에서 이들 모형은 상태 s와 행동 a에 대한 다음 상태와 보상을 제시하는 역할을 하게 된다($P(s', r|s, a)$).

강화 학습의 모형 기반 접근법에서 계획(planning)이란 개념은 모형에 주어지는 어떤 입력으로서, 그 모형화된 환경과 에이전트의 행동 정책(policy)의 상호작

용 가운데 정책 향상을 일으키는 과정을 가리킨다. 만일 학습 에이전트에게 주어진 과제가 상태 공간 속에서 목표 상태에 도달하기 위한 최적 경로를 찾는 것이라면, 계획은 바로 상태 공간에 대한 탐색 과정을 뜻하게 된다(state-space planning). 모형 기반 강화 학습의 일반적 도식에서 이 접근법은 아래와 같은 순서를 따른다.

$$\text{모형} \longrightarrow \text{모사된 경험} \xrightarrow{\text{backups}} \text{상태 값} \longrightarrow \text{정책}$$

위 도식에서 계획(planning)은 '모사된 경험'(sampled experience)과 '상태 값'(values)을 아우르는 과정으로 여길 수 있다. 그리고 가치 함수에 의해 값이 갱신된(backups) 상태 값에 의거하여 학습 에이전트는 자신의 행동을 목표에 가깝게 접근하도록 조절(control)할 수 있게 된다. 모형 기반 강화 학습의 이와 같은 도식은 아래의 모형 자유 기반 학습의 도식과 비교해보면 그 차이가 분명하게 드러난다.

$$\text{환경} \longrightarrow \text{경험} \xrightarrow{\text{backups}} \text{상태 값} \longrightarrow \text{정책}$$

위 도식에서 'planning'이라고 지칭하는 과정은 주어지는 경험 내용이 사전에 구성된 모형에 따른다는 차이만 있을 뿐 기본적인 학습방식은 거의 동일하다. 그림 5.1은 모형 기반 강화학습의 주요 개념들 간의 관계를 표현하는 도식으로, 이 그림을 통해 '모형'과 '계획', '가치 함수'간의 영향력 행사 관계를 명시적으로 이해할 수 있다. 이 그림에서 모형 학습(model learning)은 에이전트가 경험 내용을 내면화하는 과정을 가리키는데, 가치/정책 함수가 경험을 통해 직접적으로 학습이 되는 경우인 'Direct RL'(직접적 강화학습)과 대비된다. '계획'을 통한 학습은 실제 목표지점(goals)과 관련이 깊은 상태와 행위에 초점을 맞춰진 경우 학습 효율성이 크게 향상되는데, 이는 목표와 크게 관련이 없는 상태들에 대한 탐색의 비용을 줄일 수 있기 때문이다.

만일 우리가 인지적 관점을 통해 모형 기반 강화학습 접근법을 해석하면, "모형 학습 과정"을 에이전트가 경험 내용의 내적 표상을 구성 혹은 요약하고 이를 이용해 향후 학습을 보다 효율적으로 하고자 하는 시도라고 볼 수 있다. 이는 학습의 내면화를 통한 잠재 학습(latent learning)의 양상과도 유사하고, 실제로도 초창기 모형 기반 강화학습 연구들은 잠재 학습 이론(Tolman, 1948)으로부터 영향을 받았다(Singh, 1994; Agre & Chapman, 1990; Bertsekas & Tsitsiklis, 1989).

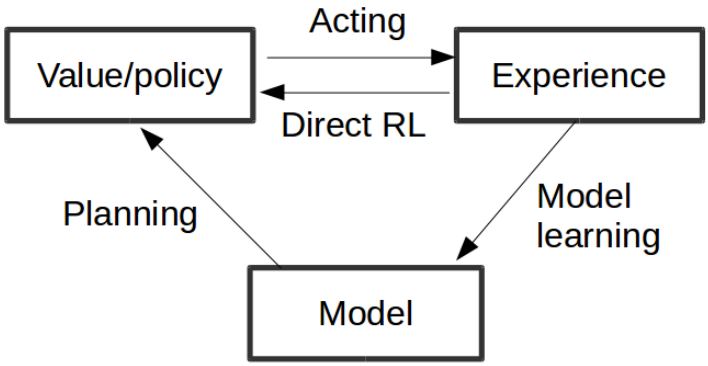


그림 5.1: 모형 기반 강화학습 주요 요소들의 관계도

하지만 모형 기반적 접근은 모형 학습에서 발생할 수 있는 오류, 즉 에이전트가 경험으로부터 부정확한 표상을 형성했을 때(wrong model), 이로 인한 학습의 정확도 감소를 유발할 수 있다. 따라서 환경에 대한 모형이 구성된 이후에 환경이 변화할 경우에, 이런 변화에 따라 모형을 수정하는 것이 어려울 수 있다. 그림 5.2은 바로 이런 경우를 잘 보여주는 연구인데(Sutton & Barto, 1998, p. 237), Dyna-Q 모형은 모형기반 강화학습 방법의 한 단순한 예제로 정책 값 업데이트가 일어날 때마다 저장된 경험으로부터 샘플링한 상태와 행위에 대한 행위값(action value)을 갱신하여 전체 학습의 효율성을 높인다. 그림 5.2의 Dyna-Q+ 모형은 탐색한지 오래된 상태들에 대한 탐색에 가중 보상값을 부여해 탐색을 진작하는 모형이며, Dyna-AC는 Q-learning 대신 Actor-Critic 모형을 사용한 모형을 가리킨다. 이 실험 결과는 Dyna-Q+ 모형과 같이 특정 탐색에 가중치를 부여하는 방식만이 이 과제에 적절하게 적응했음을 가리키며, 동시에 모든 모형이 학습 환경이 변화한 직후에 새로운 경로를 찾는데 상당한 비용을 소모했음을 보여준다.

강화 학습환경으로부터 경험 데이터를 얻는 것이 여의치 않을 경우에, 모형 기반 강화 학습은 상태 값 학습을 위한 많은 경험 내용을 스스로 얻을 수 있다는 큰 장점이 있다. 하지만 어떤 복잡한 환경을 모형화하는 것 자체가 쉽지 않은 경우, 이 접근법은 시작 단계에서 매우 큰 부담을 지게 된다. 또한 환경에 대한 모형 구축은 다양한 방식으로 이루어질 수 있고 따라서 높은 자유도가 따르는 작업이므로, 학습 환경에 대한 알고리즘 설계자의 세심한 이해가 요구된다.

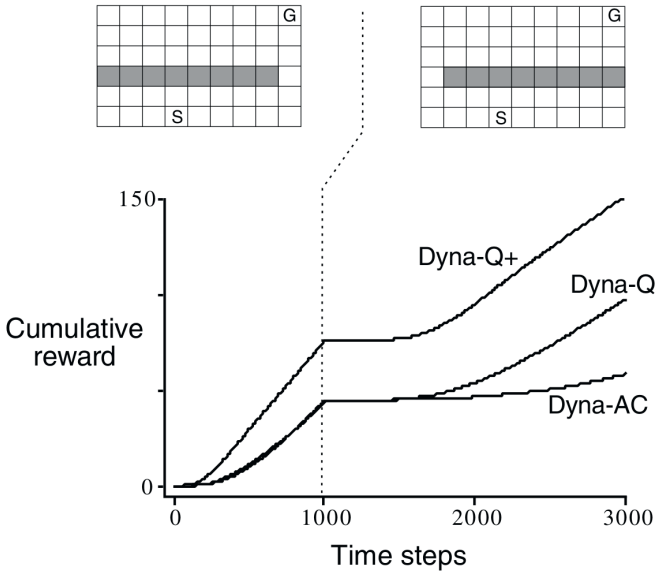


그림 5.2: 차단 벽 미로(blocking maze) 과제에서 Dyna 모형들의 성능 그래프. 위의 왼쪽 미로는 1000 번째 시간 포인트에서 오른쪽 미로로 변화한다. 에이전트는 갈색 칸들을 통과할 수 없다

5.2 함수 근사와 상태 표상

이산적(discrete) 상태-행위 공간을 활용하는 강화 학습에서 상태 공간과 행위 공간의 크기가 매우 크다면 시간-공간 계산 복잡도가 매우 커지게 되므로 실제로 쓰기가 어려운 학습 방법이 되고 만다. 따라서 패턴 인식 알고리즘을 통해 이러한 매우 많은 가능한 상태 혹은 행위를 일반화(generalization) 시켜주는 방법이 필요하다. 그리고 이 방법을 함수 근사법(function approximation method)이라고 부른다.

이 함수 근사법은 근사 함수를 이용해서 입력으로 들어온 상태가 어느 정도의 가치를 갖는지를 학습시켜 이 근사 함수를 상태 혹은 행위 가치의 측정에 사용한다. 이 문제는 기계학습의 지도 학습 전통 내에서 많이 연구되고 다양한 기법들이 개발 되었는데, 이를 통해 매우 큰 상태 공간을 가진 문제가 요구하는 기억공간이나 계산량을 크게 감소시킬 수 있다. 그리고 함수 근사법은 크게 가치 함수의 근사에 모수(parameters)를 이용하는 모수적 방법(parametric approach)과 모수를 사용하지 않는 비모수적 방법(non-parametric approach)로 나눌 수 있다.

5.2.1 모수적 방법

모수적 방법을 통한 함수 근사법은 크게 선형 근사법과 비선형 근사법으로 더 나눌 수 있는데, 선형 근사법은 미리 정의된 자질 함수 $\phi(s)$를 통해 어떤 상태 s를 자질 벡터로 표상한 후에 이 벡터를 어떤 가중치 벡터 W와의 내적을 통해 상태 가치를 계산한다. 이것을 수식으로 표현하면 아래와 같다.

$$V(s, W) = W^T \phi(s) = \sum_{i}^{d} w_i \phi_i(s)$$

그러면 확률 기울기 하강(Stochastic Gradient Descent)을 이용하여 점진적 방식으로 가중치 벡터 W를 다음과 같이 갱신할 수 있다.

$$W_{t+1} = W_t + \alpha \Big[(R_{t+1} + \gamma V(S_{t+1}, W_t)) - V(S_t, W_t) \Big] \phi(S_t)$$

상태를 자질 벡터로 사상하는 방법은 대표적으로 다항식(polynomials)을 이용하는 방법을 비롯하여, FBF(Fourier Basis Function), 타일 코딩(Tile Coding), 그리고 RBF(Radial Basis Function) 등이 흔히 쓰인다. 이러한 선형 근사법은 전역 수렴 가능성을 보장한다는 점에서 강점을 가지나, 입력 상태의 복잡한 양상을 표현하고 학습하기에는 부족하다는 한계점이 흔히 지적된다.

비선형 함수 근사법의 대표적인 방법은 인공 신경망을 활용하는 방법이 꼽힌다. 강화학습의 역사에서 인공 신경망을 성공적으로 활용한 대표적인 초창기 케이스는 Backgammon이라는 보드 게임의 상태값 근사 함수로 다층 신경망 모형을 사용한 연구이다(Tesauro, 1995). 그림 5.3 (Sutton & Barto, 1998, p. 263)은 바로 이 TD-Gammon 모형에 사용된 인공 신경망 모형을 보여주는 그림으로, 시간차 오차(temporal-difference error)는 여기서 역전파 알고리즘을 통해 신경망 학습에 이용되었음을 볼 수 있다. 그리고 가장 상위의 출력 노드의 값은 현재 게임에서 플레이어가 이길 확률을 표현하게 된다.

심층 신경망 모형(deep learning models)은 기존의 다른 패턴 인식 모형들에 비해 매우 복잡한 입력 패턴도 잘 학습할 수 있는 것으로 알려지고 있고, 최근 이와 같은 원리의 심층 인공 신경망 모형은 Atari 비디오 게임과 바둑과 같은 전통 보드

게임의 함수 근사법으로 사용되었다(Mnih et al., 2015; Silver et al., 2016). 특히 Covolutional Neural Network(CNN)과 같은 모형은 시각적 데이터에 효과적으로 적용 가능하여(LeCun, Bottou, Bengio, & Haffner, 1998) 이 연구들에서 상태에 대한 함수 근사법으로 성공적으로 이용되었다(Mnih et al., 2015; Silver et al., 2016).

5.2.2 비모수적 방법

비모수적 함수 근사법의 대표적인 방법은 기억 기반 함수 근사법(memory-based function approximation)으로, 이 방법은 어떤 상태에 대한 상태 값이 필요할 때, 기억 속에 저장된 학습 예제(training examples)를 인출하여 이들로부터 주어진 상태 값(query state)에 대한 값을 예측한다 (Atkeson, 1992; Atkeson, Moore, & Schaal, 1997; Baird & Klopf, 1993).

비모수적 방법의 이런 전략은 근사하고자 하는 가치 함수의 형태를 모수적 형식에 한정할 필요 없이, 보다 초점을 맞추고자 하는 요인이나 자질을 중심으로 상태 값을 근사할 수 있다는 장점을 가진다. 또한 양질의 학습 예제를 축적시켜 근사 정확도 향상을 유도할 수 있으며, 불필요한 상태 공간 탐색에 드는 비용을 거의 없앨 수 있다는 점에서 차원의 저주(curse of dimensionality)를 회피하는 한 방법을 제공한다.

기억 기반 접근법을 통해 주어진 상태와 가까운 상태를 찾을 때는 흔히 벡터

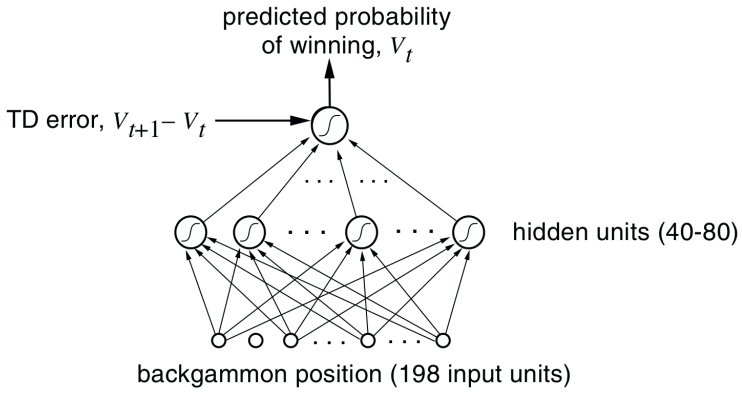


그림 5.3: TD-Gammon에 이용된 신경망 모형

공간 상에서의 거리 측정 함수(예를 들면, 코사인 거리)를 사용하여 가장 이웃한 (nearest neighboring) 상태를 찾는다. 물론 그 외에도 회귀법(regression)이나 분류 알고리즘 같은 방법이 응용될 수 있으나, 공통적인 점은 그 탐색과정은 어디까지나 일시적이며 국소적(local)이어서 완료 후 그 과정에 대한 정보는 버려진다는 점이다 (Peng, 1995; Tadepalli & Ok, 1996; Friedman, Bentley, & Finkel, 1977).

제 II 편

실험

제 6 장

시간차 기반 단어 감정 극성 값 측정법

이 장은 텍스트의 등장하는 단어의 가치를 평가하는 실험 과제를 서술한다. 그리고 여기서 말하는 가치란 텍스트의 정서적(emotional) 측면과 관련된 값을 의미한다. 자연어 처리 연구에서 특정 단어의 감정 극성 값은 히 'semantic orientation'이란 용어로 불리고, 전통적으로 이 개념은 텍스트의 감정 유형(sentiment class) 분류를 위한 어휘집(lexicon)의 한 요소로 연구되었으나(Hatzivassiloglou & McKeown, 1997; Wiebe, 2000; Taboada, Anthony, & Voll, 2006), 비지도 알고리즘을 통해 자동적으로 측정하려는 시도도 있었다(Turney, 2002; Zagibalov & Carroll, 2008). 본 장의 실험은 텍스트에 등장하는 단어들의 이 의미 방향성(sentiment orientation) 혹은 감정 극성 값(sentiment polarity) 측정을 강화 학습의 시간 차 학습(temporal-difference learning)을 통해 실시간으로, 즉 온라인(on-line) 방식으로 측정하는 구체적인 방법을 기술한다.

6.1 실험의 기본적 설계

본 실험에서 단어의 극성 값 측정 문제는 시간차 학습 방법을 통한 단어의 가치 예측의 문제로 간주되었다. 한 텍스트 혹은 한 문장은 단어의 연쇄열로 여길 수 잇으므로, 어떤 텍스트는 시작점과 종결점을 가지는 하나의 일화(episode)로서 각 단어들의 점증적(incremental)으로 덧붙여지는 단어열들로 이루어진다. 따라서 어떤 텍스트가 갖는 단어 수가 T 라면 그 텍스트는 단어의 연쇄열 $w_1, w_2, \ldots, w_t, \ldots, w_T$ 이며 각 단어는 시간 시점 t 에 따라 하나씩 순차적으로

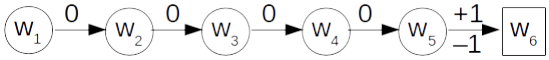


그림 6.1: 6개 상태로 이루어진 마르코프 보상 과정의 예제. 각 상태는 왼쪽에서 오른쪽으로 1의 확률로 전이된다. 각 화살표 부근에 표시된 숫자는 이행 과정에서 주어지는 보상 값을 가리키고, 숫자가 둘인 경우엔 보상 값의 경우가 두 가지임을 의미한다.

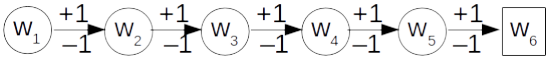


그림 6.2: 6개 상태로 이루어진 마르코프 보상 과정의 예제 2. 이 마르코프 보상 과정에서는 상태 전이 때마다 보상 값이 $+1$ 혹은 -1이 주어진다고 가정된다.

나타난다고 보았다. 이 과정은 사람이 문장을 왼쪽에서부터 하나씩 차례로 읽어나가는 과정과 매우 유사하다. 그리고 사람은 텍스트를 순차적으로 읽어나갈 때 그 내용에 대한 즉각적 이해를 수행하므로 각 단어가 추가될 때마다 그 내용에 대한 보상 값 $r \in \mathbb{R}$이 주어진다고 가정하였다. 그림 6.1과 그림 6.2는 본 실험에서 사용된 마르코프 보상 과정을 6개 상태의 연쇄로 표현한 예제로, 둘의 차이는 모든 상태 전이시 분류 레이블에 따른 보상값이 주어지는지의 여부 뿐이다.

6.1.1 상태

상태는 전체 말뭉치의 어휘 중 하나로 정의 되었다. 그러나 물론 여기서 단어를 한 단위로 정한 상태 정의는 임의적이며, 그것은 음소와 같이 더 작거나 구(phrase)와 같이 더 큰 단위로 사전에 결정될 수 있다. 다만 본 실험에서는 단어의 감정 값을 측정하는데 관심을 두었으므로 그 단위를 단어로 한정하였는데, 각 문장의 단어들은 공백을 기준으로 분리되었으며 각 단어는 고유한 한 상태와 사상되었다. 그러므로 어휘 크기의 배열이 측정된 상태 값을 저장하기 위해 사용되었다.

6.1.2 보상

보상 r은 종결 상태에 도달하기 전까지 상태 이행 때마다 주어진다. 일반적인 감정 분석 과제에서 텍스트에 대한 감정 태그는 긍정(positive) 혹은 부정(negative) 중 하나인데, 본 실험에서는 긍정 태그는 실수 값 $+1$로, 부정 태그는 -1로 변환하여 보상 값으로 정의하였다. 만일 한 텍스트 내에서 단어가 연속적으로 관찰 가능할 때마다 감정 태그가 주어지지 않는다면, 보상 값으로 0이 항상 주어진다고 가정

하였다.

6.2 실험 1

이 실험은 시간 차 학습을 이용해 측정한 감정 극성 값에 따라 자질 선택(feature selection)을 함으로써 그 선택의 결과가 유의미함을 보이기 위한 것이다. 단어들을 자질로 이용하는 종류의 감정 분석 과제에서는 전체 어휘를 자질로 쓰는 것 보다 정보 가치가 높은 N개 단어들을 선별하여 사용하는 것이 일반적이다. 이 과정에서 특정 N개 단어를 선택하는데 기존의 다양한 선택 알고리즘들이 존재한다(Yang & Pedersen, 1997; Sharma & Dey, 2012; Uğuz, 2011). 본 실험의 목적은 시간 차 알고리즘을 이용한 자질 선택 과제의 수행 결과가 기존의 전통적 기계 학습 내에서의 자질 선택 방법과 견줄만 하다는 것을 보이고자 하는 것이다. 그리고 이를 통해 시간 차 학습에 따라 측정된 감정 극성 값의 유의미성은 간접적으로 검증된다고 본다.

6.2.1 데이터

Polarity Dataset v2.0 1,000개의 긍정 영화평과 1,000개의 부정 영화평 자료(Pang & Lee, 2004)이며 전처리가 되어 있다.

6.2.2 비교 조건

Document Frequency 어떤 단어가 텍스트 데이터 셋 전체에서 등장한 문서의 수를 가리킨다. 따라서 이 문서 빈도가 큰 단어는 데이터 상에서 고루 분포 되어 있다고 볼 수 있다. 본 연구에서는 상위 10,000개의 문서빈도를 가진 어휘들을 자질로 선택하였다.

Information Gain IG는 기계 학습 기반 분류 과제에서 현재까지도 널리 이용되는 자질 선택 방법 중 하나로, 텍스트의 특정 단어 자질이 존재하는지 아닌지에 따른 텍스트 분류 과제상 정보 이득을 비트 단위로 측정한다. 어떤 문서 집합 D에

대한 분류 태그를 예측하는데 필요한 정보는 아래와 같이 정의된다.

$$Info(D) = -\sum_{i=1}^{m}(P_i)\log_2(P_i)$$

여기서 m은 분류 범주의 수를 가리키고, P_i는 문서 집합 D에 속한 임의의 문서 d가 분류 범주 C_i에 속할 확률을 가리킨다. 자질 집합 $A = \{a_1, \cdots, a_v\}$가 있을 때, 문서 집합 D는 이 자질 공간에 따라 $\{D_1, \cdots, D_v\}$로 분리될 수 있다. 그럼 아래와 같이 자질 공간에 대한 정보를 정의할 수 있다. 여기서 $\frac{|D_j|}{|D|}$는 D의 j번째 분할(partition)이 갖는 가중치를 의미한다.

$$Info_A(D) = -\sum_{j=1}^{v}\frac{|D_j|}{|D|} \times Info(D_j)$$

아래는 자질 A에 대한 최종적인 수식이다.

$$IG(A) = Info(D) - Info_A(D)$$

CHI statistic χ^2 statistic (이하 CHI)는 텍스트의 분류 범주와 단어 자질 사이의 관련도를 측정하는데 있어 단어의 출현 기대 분포로부터의 이탈 정도를 측정하여 그 정보가를 표현한다. 그리고 여기서 어떤 단어의 출현 가능성은 다른 단어에 대해 상호 독립이라고 기본적으로 가정한다. 그 수식은 아래와 같다.

$$CHI(t,c) = \frac{N \times (AD - CB)^2}{(A+B) \times (B+D) \times (A+B) \times (C+D)}$$

위 수식에서 c는 분류 범주를, t는 단어 어휘를, N은 전체 문서 수를, A는 t와 c의 공출현(co-occurrence) 빈도를 가리키며, D는 c 혹은 t 중 어느 쪽도 등장하지 않은 횟수를 뜻한다. 그리고 C는 c가 t와 D 없이 등장했던 빈도를, 마지막으로 B는 c 혹은 t 중 어느 쪽도 나타나지 않았던 경우의 빈도를 가리킨다.

TF-IDF 이 방법은 단어의 빈도와 그 단어에 대한 상대 문서 빈도의 역수를 곱하여 단어의 중요도를 측정하는 전통적인 방법인 TF-IDF 방법을 이용한다. 각

문서의 단어에 대해 TF-IDF 값을 구하고 전체 문서 집합에 대해 그 평균을 구한 값을 이용해 단어들을 정렬하였다. 그리고 위 다른 방법들과 마찬가지로 상위 10,000개 값을 가진 단어 자질들을 선택하여 사용하였다. TF-IDF 수식은 아래와 같다. $TF(t)$는 단어 t가 한 텍스트에서 등장한 단순 빈도를 가리키며 N은 전체 문서의 수를 가리킨다.

$$TF\text{-}IDF = TF(t) \times \log_{10} \frac{N}{t \in d : d \in D}$$

Simple Averages 본 실험의 방법 중 가장 단순한 자질 선택 방법으로 어떤 어휘 t가 긍정 텍스트에서 나타났을 경우 $+1$을, 부정 텍스트에서 출현했을 때는 -1을 더하고 그 합에 대한 평균을 구해서 해당 어휘의 자질 값으로 삼았다. 위 기계학습 전통에서 나온 자질 정보가 계산법과는 달리 이 방법은 극성 값의 형태를 띠기 때문에 상위 5,000개와 하위 5,000개의 어휘를 선택하여 사용하였다.

6.2.3 실험 방법

본 실험은 시간 차 학습의 대표적인 알고리즘인 TD(λ) 모형(Sutton, 1984, 1988)과 이 모형의 on-line 측정 기제를 이론적으로 완성한 모형인 True On-line TD(λ)를 사용하였다(Seijen & Sutton, 2014; Van Seijen et al., 2016). TD(λ) 모형은 보상 감쇄율인 γ 외에도 학습률(learning rate)과 흔적 감쇄율(trace-decay rate) 모수를 적절하게 결정하는 것이 요구되므로, 먼저 이 모수 공간을 탐색하기 위한 실험을 수행하였다. 학습률은 0.1에서 0.5까지 0.1씩 값을 증가시켰으며 흔적 감쇄율은 0.0에서 1.0까지 역시 0.1씩 증가시키며 결과를 관찰하였다.

실험 수행도는 10-fold cross validation 테스트를 통해 그 평균 정확도를 보았으며, 학습 데이터로부터 단어들의 감정 극성 값을 계산한 후에 최고 긍정가를 갖는 단어 5,000개와 최고 부정가를 갖는 단어 5,000개를 골라 단어 자질로 선택하였다. 그 후에 나이브 베이즈 분류기(Naive Bayes classifier)에 선별된 단어 자질 집합을 주어 학습을 시킨 후에 테스트 데이터에 대해 그 정확도를 기록하였다. 따라서 한 번의 실험에서 전부 10개의 정확도가 나오며 이것들의 평균이 바로 그 조건의 수행도로 기록되었다.

Significance traces 섹션 **4.4**에서 기술된 적격 흔적도(eligibility traces)는 어떤 상태 s의 방문 빈도와 방문이 이루어진 시점의 최근성(recency)에 대한 가중치 모형이라고 생각할 수 있다. 즉, 이 방법을 따르면 최근에 그리고 자주 등장하면 더 많은 가중치를 받게 된다. 하지만 자연어 텍스트에서 불용어 혹은 불용어는 아니지만 텍스트의 토픽이나 도메인에 의해 자주 등장하는 고빈도의 단어들이 흔히 존재한다. TD(λ) 알고리즘에서 이들 단어들의 적격 흔적도는 매우 심하게 오르락 내리락 하는 패턴을 갖게 되는데, 이들 단어들의 실제 의미적 기여는 매우 작다는 점을 볼 때, 학습 과정에서 예기치 않은 오차 요인으로 작용할 위험성이 있다. 특히 짧은 시간 안에 상태가 많이 나타날수록 누적적으로 적격 흔적도가 증가하는 흔적도 유형들에선(accumulation traces, Dutch traces) 더욱 그러하다. 이런 경우를 예방하기 위해 본 연구에서는 수식 6.1를 통해 단어의 상대 문서 빈도의 역수를 적격 흔적도에 반영하는 조건을 준비하였다. 일반적인 inverse document frequency와는 다르게 로그 값을 사용하지 않는다는 점을 유의하라.

$$idf(w) = \frac{N}{df(w)} \tag{6.1}$$

이 상대 문서 빈도는 어떤 단어 w의 등장이 잦아서 모든 문서에서 관찰 된다면 1이 되고, 그 출현 빈도가 매우 작다면 그 값은 매우 커지게 된다. 본 연구는 이 점을 이용하여 어떤 단어의 출현 빈도(frequency)와 최근성(recency)을 아울러 그 현저성(saliency)을 반영하고자 하였다. 수식 6.2은 이런 맥락에서 siginificance traces라고 여기서 명명 되었다.

$$e(w) \leftarrow \tanh \frac{idf_{incr}(w)}{4} \tag{6.2}$$

위 수식에서 idf_{incr}은 이 값이 점진적 방식으로 계산된다는 것을 가리키고 있다. 하지만 물론 미리 추정치를 계산 후에 사전 형태로 저장하여 사용하는 방법도 가능하나, 본 연구에서 시도되지는 않았다. 분자 위치의 양수 '4'는 idf 값이 0.5일 때, 즉 어떤 단어가 한 텍스트에 출현할 가능성이 50%에 가까울 때, 최종적인 값이 0.5에 가까운 값인 0.46이 되도록 만들어 주는 스케일링 상수 역할을 한다.

tangent-hyperbolic 함수의 인자는 언제나 양수이므로 이 흔적도는 기본적으로 $[0, 1]$의 범위를 가지나 idf는 최소값이 1 이하일 수는 없으므로 이 흔적도의 치역은 $[0.24, 1]$이 된다($\tanh 1/4 \approx 0.24$). 이에 따르면 어떤 단어가 매우 드물게 등장한 단어라면 이 흔적도는 1에 가깝게 된다. 따라서 이 적격 흔적도는 고빈도의 단어들의 흔적도를 학습 기간 내내 낮게 유지되도록 만드는 기능을 하게 된다.

학습의 세부 사항 학습은 섹션 **4.4**의 알고리즘 2를 따라 수행되었다. 적격 흔적도가 비현실적으로 증가하거나 감소하는 경우를 방지하고자 흔적도가 $[-100, 100]$을 넘지 못하도록 하였으며, 단어의 상태 값을 학습 전체 기간에 대해 계산하기 위해 다음의 공식, $\mu_k = \mu_{k-1} + \frac{1}{k}(x_k - \mu_{k-1})$을 이용하여 그 평균을 상태 값 갱신 때마다 계산하였다. 여기서 x_k는 k번째 보상 값(+1, −1 혹은 0)을 가리킨다. 그리하여 학습이 끝나면 최종적으로 전체 시간 시점들에 걸친 상태 값들의 평균이 주어지게 된다. 이런 방법을 취한 이유는 이 실험에서는 단어의 특정 시점에 대한 감정 극성 값이 아니라 학습 데이터 전체에 대한 값이 필요하기 때문이다.

6.2.4 결과

그림 6.3은 모수 탐색 실험의 결과를 보여준다. 이 결과에 따르면, 가장 높은 정확도는 모두 흔적 감쇄율이 1에 도달 했을 때 나왔다. 따라서 TD(1)이 이 실험 환경에서 가장 효과적인 모형인 것으로 나타났다. 전반적으로 흔적 감쇄율이 낮아질수록 그 성능도 하락하는 패턴을 보였으며, 흔적도가 누적적으로 증가하는 유형의 적격 흔적도를 이용한 조건은 학습률이 증가함에 따라 상대적으로 심한 성능 하락 현상을 보였다. 그에 비해 significance 흔적도 조건은 학습률의 증가에도 심한 성능 하락을 보이지 않았다. 그림 6.3의 결과로부터 각 TD 모형 조건별로 최적의 모수 값은 테이블 6.1에 보기 쉽게 제시되어 있다.

테이블 6.2은 실험 1의 결과를 요약하여 보여주는데, 이 결과는 테이블 6.1에 제시된 최적 모수 값들을 사용한 결과이다. 이 결과는 TD 방법들과 Information Gain이 가장 효과적인 어휘 자질 선택을 하였음을 잘 보여주고 있다. 가장 높은 성능을 기록한 모형은 accumulation trace를 사용한 TD 방법으로 정확도 0.84를 기록하였다. 그리고 True Online TD 방법을 제외한 모든 TD 방법들은 0.83의

Traces	λ	α
Accumulation	1	0.1
Replacing	1	0.3
Dutch	1	0.2
Significance	1	0.3
True Online TD	1	0.1

표 6.1: TD 방법의 최적 학습률과 흔적 감쇄율

Method	NB Accuracy
TD(1) with accumulation	**0.84**
TD(1) with replacing	0.83
TD(1) with Dutch	0.83
TD(1) with significance	0.83
True Online TD(1)	0.78
Simple Averages	0.64
Document Frequency	0.67
TF-IDF	0.69
χ^2 statistic (CHI)	0.66
Information Gain	0.83

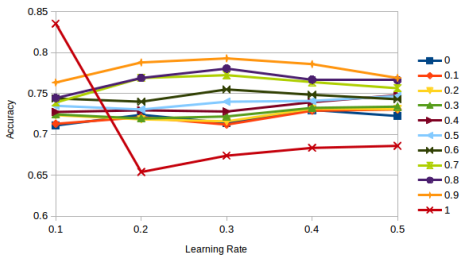
표 6.2: TD 방법과 그 외 자질 선택 방법들에 따른 감정 분류 결과의 정확도

정확도를 기록했다. 그러나 accumulation trace를 사용한 TD 방법과 True Online TD 방법은 학습률이 0.1을 벗어나면 심하게 성능이 하락하는 양상을 보였는데, 이는 나머지 TD 방법들의 최적 학습률이 0.2 혹은 0.3을 기록한 것과 대조적이다.

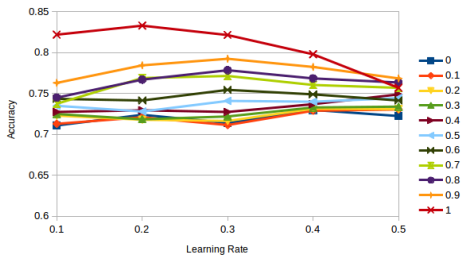
결과 중에서 단순 평균값을 사용한 Simple Averages 조건이 가장 낮은 성능을 기록했으며, Information Gain을 제외한 다른 자질 선택 방법들은 모두 0.7의 정확도 이상을 기록하지 못했다. Information Gain은 0.83의 수행능력을 기록하며 TD 방법들의 결과와 크게 다르지 않은 결과를 보였다. 이 모든 결과는 TD 방법들이 어휘 자질 선택에 있어 기존에 널리 사용되는 선택 알고리즘들을 앞서가나 적어도 뒤지지 않음을 보여주는 것이다.

6.3 실험 2

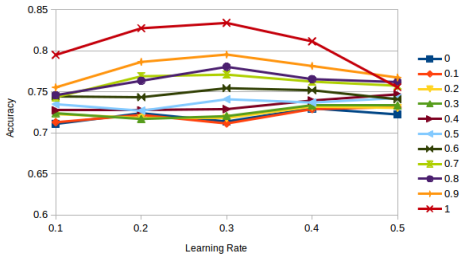
실험 1은 TD 방법을 통한 감정 극성값 측정이 효과적임을 보여준다고 할 수 있으나, 이는 간접적인 방법에 의존한 것이었다. 이는 일반적인 감정 분석을 위한 데이터셋이 텍스트에 대한 감정 분류 정보를 제공하지, 개별 어휘에 대해서는



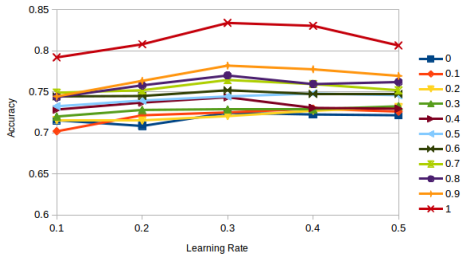
(a) Accumulation traces



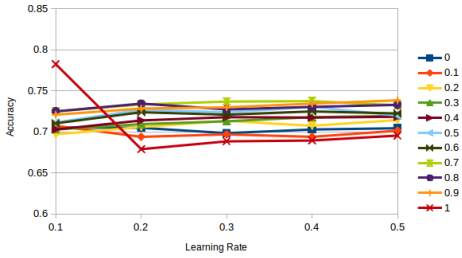
(b) Dutch traces



(c) Replacing traces



(d) Significance traces



(e) True Online TD

그림 6.3: TD 방법의 자질 선별 결과에 따른 감정 분류 정확도.
x 축은 학습률을 가리키고 각 선들은 흔적 감쇄율을 뜻한다.

그런 정보를 제공하지 않는 제약으로부터 비롯된 것이다. 실험 2에서는 각 단어 어휘들에 대해 인간 주석자가 감정가를 직접 주석한 데이터셋을 가지고 보다 직접적인 비교를 수행하였다.

6.3.1 데이터

Stanford Sentiment Treebank 이 데이터는 11,855개의 영화평 문장들의 모든 개별 어휘들에 대한 감정 극성 주석값을 포함하고 있다 (Socher et al., 2013). 각 주석은 0에서 1 사이의 실수 값이며 이는 세 명의 인간 주석가들의 평점에 따른 것이다. 본 실험에서는 총 두 개의 어휘 집합을 구축하여 사용하였다. 먼저 데이터의 총 21,684개 어휘 셋을 정하였고, 품사 태깅을 통해 품사 별로 단어들의 감정 극성 값의 분산이 가장 컸던 세 개 형태소(형용사, 형용사 최고급, 부사 비교급)를 가진 어휘들을 따로 선별하여 4,532개 어휘 셋을 만들었다. 이는 분명한 감정 극성값을 가지는 어휘들을 따로 선택하여 비교하기 위함이다.

6.3.2 비교 조건

TD 방법을 제외한 실험 1의 자질 선택 알고리즘들은 대부분 감정 극성 값을 측정하는 데는 적절치 않은 방법들이므로, 본 실험에서는 베이즈 확률을 이용하여 단어의 감정 극성값을 측정하는 방법(Potts, 2010) 을 비교 조건으로 사용하였다. 이 방법에 따라, 먼저 각 감정 분류 범주 c에 대한 빈도를 아래와 같이 구한다. 여기서 V는 어휘 집합을 가리킨다.

$$Count(c) = \sum_{t \in V} Count(t, c)$$

그리고 감정 범주 $c \in C$가 주어졌을 때, 어떤 단어 어휘 t가 그에 속할 확률을 아래의 수식을 이용해 구한다.

$$Pr(t|c) = \frac{Count(t, c)}{Count(c)}$$

Method	Pearson (full)	Spearman (full)	Pearson (reduced)	Spearman (reduced)
Bayes Prob. (Potts, 2010)	0.21	0.2	0.32	0.3
TD(1) with replacing	0.24	0.21	0.38	0.35
TD(1) with Dutch	0.24	0.21	0.38	0.35
TD(1) with significance	0.24	0.21	0.38	0.35
TD(1) with accumulation	0.24	0.21	0.38	0.34

표 6.3: 주석된 단어 극성값과 알고리즘을 통한 측정값과의 상관관계. 여기서 full set은 전체 어휘 셋을 가리키고 reduced set은 따로 선별하여 구축한 4,532개 단어 어휘 셋을 가리킨다.

그럼 아래의 수식과 같이 어떤 단어가 어떤 감정 분류에 속할지에 대한 베이즈 확률을 계산할 수 있다. 본 실험에서는 각 단어가 긍정 범주에 속할 확률을 계산하여 활용하였다.

$$Pr(c|t) = \frac{Pr(t|c)}{\sum_{c' \in C} Pr(t|c')}$$

6.3.3 실험 방법

실험 1에서 기술된 된 것과 동일한 방법으로 TD 기반 감정 극성값을 측정하였다. 흔적 감쇄율은 모두 1을 사용하였고, 학습률 역시 실험 1의 결과를 참고하여 0.1과 0.2 중 하나를 택일하여 사용하였다. 감정 극성 측정값의 정확성은 상관관계 계수를 이용하였고, Pearson 상관계수와 Spearman 상관계수 모두를 구하여 비교하였다.

6.3.4 결과

실험 2의 결과는 테이블 6.3에 제시되었다. 실험 1에서 사용된 True Online TD 조건은 이전 실험에서의 낮은 성능으로 인해 본 실험에서는 사용되지 않았다.

결과에서 볼 수 있듯이, TD 기반 측정값들은 모두 베이즈 확률값보다 주석된 값들에 대해 상대적으로 높은 상관관계를 보였다. 그리고 TD 방법과 베이즈 기반 방법 모두 선별된 어휘 셋에 대해 더 높은 상관관계를 기록했다. TD 방법들의 성능은 거의 동일한 것으로 나타났고, 따라서 적격 흔적도의 유형보다는 흔적 감쇄율과 학습률이 성능에 미치는 영향이 더 큰 것으로 보인다.

6.4 실험 3

실험 1과 2는 TD 방법을 통한 감정 극성값 측정의 효과성을 간접적, 직접적 방법에 의해 검증하였다. 이 두 실험의 TD 방법은 모두 그림 6.1을 따라 실제 보상값은 텍스트 말미에 주어진다고 가정하였다. 실험 3은 그림 6.2를 따라 새로운 단어가 시점 t에서 주어질 때마다, 문자열 $w_1, w_2, \ldots w_t$에 대한 감정 분류에 따른 보상 r이 주어진다는 실험 설계를 따른다. 하지만 이런 구성에 맞춰 정답이 주석된 데이터 셋이 부재하므로 재귀 신경망 모형을 이용한 감정 분류기(LSTM classifier)를 이용해 보상값을 할당하였다. 이 실험 구성은 텍스트의 글자 수 만큼의 분류 작업을 요하므로 계산비용의 측면에서는 비효율적이나 인간의 언어처리가 본래 점증적 처리(본 논문 2장을 참조)에 가깝다는 점에서 실제 인지처리 과정과 보다 유사하다고 할 수 있다. 또한 이 구성은 텍스트 상태의 보상값을 미리 학습시킨 어떤 모형에 의존한다는 점에서 모형 기반적 강화학습을 이용한다고 볼 수 있다. 그리고 학습된 모형에 의해 보상값이 부여되고 그 모형은 부정확한 분류를 할 수 있기 때문에, 이에 기반한 단어의 감정가 예측도 부정확할 가능성은 커지게 된다.

6.4.1 데이터

Polarity Dataset v2.0 실험 1에서 사용된 영화 평 자료(Pang & Lee, 2004)가 TD 방법의 테스트 데이터 셋으로 사용되었다. 여러 문장으로 이루어진 영화평은 문장 단위로 분리된 후에 각 문장별로, 그리고 누적적으로 증가하는 각 단어 연쇄열들에 대해 LSTM 분류기가 긍정 혹은 부정 태그를 부여한 데이터 셋이 최종적으로 TD 방법의 데이터로 이용되었다.

Large Movie Review Dataset 긍정/부정 감정 태그를 가진 25,000개 학습 데이터 셋과 25,000개 테스트 셋이 LSTM 분류기의 학습과 테스트를 위해 사용되었다. 학습 셋의 10%는 검증 데이터 셋으로 이용되었다 (Maas et al., 2011).

6.4.2 비교 조건

실험 1을 통해 가장 효과적인 자질 선택 방법으로 드러난 Information Gain 방법

이 비교조건으로 선택되었다. 이 조건에서 분석의 단위는 LSTM 분류기가 감정 태그를 부여한 한 문장이 된다.

6.4.3 실험 방법

Polarity Dataset의 모든 영화평은 문장별로 분리되고 한 문장에 대해 누적적으로 증가하는 각 단어 연쇄열에 대해 LSTM 감정 분류기가 긍정 혹은 부정 감정 태그를 부여하였다. 여기에 사용된 TD 방법은 실험 1과 2와 동일하나, 각 단어의 시간차 값들에 대한 평균을 사용하지는 않고 최종 단계의 시간차 값만을 단어 갑정가 예측치로 이용하였다. 또한 전체 데이터 셋으로부터 8대 2의 비율로 학습 셋과 테스트 셋을 10회 샘플링하여 실험한 후에 이 정확도들의 평균을 최종 성능에 대한 지표로 사용하였다.

LSTM 분류기 학습 학습 단계에서 검증용 데이터셋에 대해 학습 모수들은 튜닝되었다. 최종 모형의 LSTM 층위 수는 2로 조절되었고, 단어 임베딩 차원 수는 400으로, LSTM cell 크기는 100개였으며 Dropout 층위가 존재하여 그 모수는 0.5였다. 학습의 미니 배치 크기는 500이었으며 30번의 순회(iteration)을 거쳐 학습이 이루어졌는데, 학습률은 0.001이었고 최대 연쇄열 크기는 200으로 한정되었다. 학습 결과, 테스트 데이터 셋에 대한 최종적인 분류 정확도는 0.84를 기록하였다.

6.4.4 결과

감정 분류 정확도에 대한 본 실험의 결과는 표 6.4에 제시되었다. 이 표의 결과에 사용된 흔적 감쇄율(trace-decay rate)은 실험 1과 2에 사용된 값과 상당히 다른데 본 실험에서 최종적으로 사용된 값은 0.7이었다. 학습률은 모두 이전 실험들과 다르지 않은 값인 0.2가 사용되었다. 흔적 감쇄율 값으로 0.7이 선택된 이유는 본 실험에 대한 최적 모수 탐색 실험 결과, 실험 1과 2와는 달리 최적의 흔적 감쇄율이 1 보다 작은 것으로 나타났기 때문이다. 그림 6.4은 replacing trace를 사용한 TD 방법의 수행을 흔적 감쇄율의 변화에 따라 기록한 것이다. 그리고 이와 같은 패턴은 모든 TD 조건에서 거의 유사하게 나타났다.

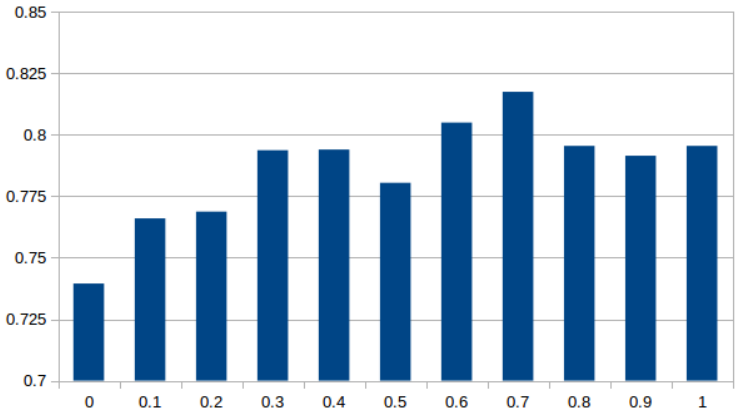


그림 6.4: 흔적 감쇄율에 따른 TD 방법(replacing traces)의 성능 그래프

Method	NB Accuracy
TD(λ) with accumulation traces	0.82
TD(λ) with Dutch traces	0.82
TD(λ) with replacing traces	0.82
TD(λ) with significance traces	0.82
Information Gain	0.75

표 6.4: LSTM 기반 보상 모형에 의존한 TD 방법과 IG 방법의 성능표

표 6.4에서 볼 수 있듯이, TD 방법들 간의 성능 차이는 없었고 비교조건인 IG 방법에 비해 높은 성능을 기록하였다(0.82 vs 0.75). 모든 조건은 실험 1의 결과에 비하면 정확도가 감소하였는데 이는 LSTM 분류기의 성능이 완전하지 않다는 것을 감안하면 예상 가능한 현상이다. 하지만 IG 방법의 하락폭은 실험 1에서의 성능 0.83에 비해 하락폭이 상대적으로 매우 크게 나타났다. 이에 비해 TD 방법들의 하락폭은 별로 크지 않은 것으로 드러났다.

그림 6.4은 TD 방법의 최적 흔적 감쇄율은 0과 1 사이의 중간 지점에 위치한다는 점을 보여 주는데, 이 차이는 실험 1과 2와는 달리 본 실험의 학습 환경의 보상체계가 달라진 데서 연유한 것으로 보인다. 본 실험의 환경에서 실제 보상값은 단어가 추가될 때마다 주어지게 되므로 이 보상값과 이에 따른 시간차 오차(TD error)에 대한 가중치 배분(credit assignment)이 최대로 길게 유지될 필요가 없게 된다. 즉, 이 가중치 배분의 범위를 최소와 최대 사이에서 적당한 위치로 정할 때, 가장 예측력이 높은 상태 값 측정을 할 수 있다는 것을 의미한다.

제 7 장

간호 진술문의 온라인 상태 값 측정법

본 논문의 서두에서 밝힌 바 있듯이, 본 연구의 초점은 강화 학습 방법론을 통해 자연어 텍스트를 점진적(incremental), 연속적(continuous)으로 처리하는 방법을 연구하는 데 있다. 그리고 앞 장의 실험에서 단어들의 연속적 상태 값을 시간차 학습을 통해서 측정하는 법을 보았다. 이 장의 실험은 반면에 텍스트의 단어가 아니라 텍스트의 문장의 상태 값을 측정하는데 목적을 둔다. 그리고 실험의 대상으로 의료 도메인의 성격을 띤 간호 진술문을 택한 데는 다음의 이유들이 있다. 첫째, 간호 진술문은 입원 환자에 대한 로그 기록과 같은 성격의 자료이기에 짧은 시간에도 매우 긴 텍스트 자료가 생성된다. 그러므로 일반적인 지도 학습 방법에 비해 시간차 학습의 적용이 계산적으로 보다 경제적이다. 둘째, 점진적(incremental) 학습 방식이 보다 알맞은 대상이다. 입원 환자의 상태는 매우 역동적이고 급변할 가능성이 항상 존재하므로, 미리 고정된 자료를 통해 학습하고 이 학습 결과를 계속해서 활용하는 것보다는 변화하는 상황 안에서 계속적으로 학습을 수행하는 강화 학습 방식이 보다 적절하다. 셋째, 일반적인 기계 학습 방법이 잘 작동하지 못할 가능성이 크다. 왜냐하면 위 도메인의 특성상 서로 상충적인 자질들이(e.g. '발열 없음', '발열 있음') 시간대만 다르게 여럿 분포하고 있기 때문에, 시간적(temporal) 요소를 반영하지 않는 학습 모형들은 불리할 가능성이 크다.

7.1 실험 목적

본 실험의 목적은 주어진 간호 진술문을 분석하여 간호 진술문이 매번 주어질

Time	Sentence
20120627055500	심호흡 교육함
20120627055500	구강간호 시행함
20120627063000	오심 감소함
20120627063000	침상안정 중임
20120627080000	수액 주입중임Right.arm 20G
20120627080000	정맥주사부위 통증,부종,발적 없음
20120627080000	일혈 위험약물과 증상에 대해 교육함
20120627080000	정맥주사부위 통증,부종,발적 없음
20120627080000	금식 중임
20120627080000	갈증 없음
20120627080000	수분부족 증상 관찰함

표 7.1: 간호 진술문 데이터 예제

때마다 그 진술문의 상태가 약물 유해 반응(Adverse Drug Reaction: ADR)과 얼마나 관련이 있는지, 즉 환자가 약물 유해 반응에 빠질 위험성이 얼마나 되는지를 실시간으로 모니터링 하는 방법을 연구하는 것이다.

약물 유해 반응, 이른바 ADR은 질병의 예방, 진단 및 치료 목적으로 적절한 투여 경로를 통해 사용량의 약물을 투여하였을 때 의도하지 않게 발생한 새로운 반응(WHO)을 가리키는데, 미국의 입원 환자 8백만명 이상을 대상으로 조사한 연구에 따르면 입원 환자의 약 1.73%는 한 가지 이상의 ADR을 경험한 것으로 나타났다(Bond & Raehl, 2006). 따라서 본 실험은 병원에 입원한 환자들의 간호 진술문에 의존하여 약물 부작용 경고 시스템에 필요한 텍스트의 상태 값을 온라인 방식으로 측정하는 것을 목표로 한다.

7.2 데이터

본 연구의 데이터는 1994년 6월 1일부터 2015년 7월 8일까지 아주 대학교 대학병원 입원 환자 중 약물 부작용 보고가 데이터에 한 번 이상 있었던 환자 4,158명과 약물 부작용 보고가 전혀 없었던 4,158명의 환자들의 간호 진술문 집합이다. 표 7.1은 이 데이터가 어떠한 성격의 자료인지를 보여주는 예제이다.

본 데이터의 진술문은 간호사들이 진술문의 템플릿을 선택하고 직접 편집하여 입력하는 형태로, 변이가 매우 다양하여 문장 유형의 전체 수는 837,293개 였다. 장기 입원한 환자의 경우 퇴원시까지 받게 되는 진술문의 수는 매우 커서 전체

Sentence	Label
오심 없음	
구토 없음	
침상안정 중임	
2(구두처방)Tramadol	T
주관적 진술:"머리랑 배가 아파요."	T
복부 통증 호소함	T
통증 양상 관찰함	T
통증완화 위한 다양한 방법 격려함	T
의사에게 알림	T
심호흡 교육함	
구강간호 시행함	
오심 감소함	

표 7.2: 약물 유해반응과 관련 있는 문장 주석의 예

데이터 셋에서 가장 많은 경우는 10,625개 진술문을 기록했다. 한 개인이 받는 진술문 갯수의 평균은 421개였다.

7.3 실험 방법

본 실험은 시간차 학습 알고리즘인 TD(λ)를 이용해 간호 진술문의 온라인 상태를 측정하고자 시도하였다. 이를 위해 비모수적 함수 근사법인 기억 기반 함수 근사를 통해 각 진술문을 전부 7개 중 하나의 상태로 사상하고 그 상태 값은 시간차 학습에 따라 계산하여 구하였다.

7.3.1 상태

한 간호 진술문은 그것이 의미하는 기능적 사건 공간의 한 사건(functional event)에 속한다고 정의되었다. 그리고 자연어 문장의 전체 집합은 거의 무한하고 이를 이산적 상태 공간에서 구축하는 것은 사실상 불가능하기 때문에 함수 근사법을 이용하여 상태표상 및 가치함수를 정의하는 방법이 고려되었다. 본 실험에서는 비모수적 방법인 기억 기반 방법론을 취하여 한 명의 간호 진술문 전문가가 ADR과 관련 있는 것으로 보이는 문장들을 미리 표시하고 이를 진술문 상태 표상을 위한 학습자료로 이용하고자 하였다. 표 7.2는 이런 문장들의 예를 보여준다.

무작위로 샘플링된 400명 ADR 환자 데이터에 대해 간호 데이터 전문가가

State	Event	Example
0	Unknown event	CT찍고 옴.
1	Drug related event	epocelin 1g 투여 함
2	Abnormal reactions	피부 가려움 호소함부위:both arm
3	Doctor related event	의사에게 알림의사:xxx
4	Subjective response	주관적 진술:몸이 가벼워요
5	1 & 2 event	Tramadol 맞은 후 구토 2회 함
6	4 & 2 event	주관적 진술:속이 계속 울렁거리네요.

표 7.3: 간호 진술문의 상태 공간 지시 기호와 그 의미

주석한 자료로부터 347개의 고유 진술문을 획득하였고, 이 문장들이 의미적으로 가리키는 사건 범주에 따라 표 7.3와 같이 그 사건 유형을 분류하였다. 그리고 347개 ADR 유관 문장에 195개 ADR 무관 문장(사건 범주 0)을 뽑아 더하여 어떤 간호 진술문을 위 범주 중 하나로 사상할 수 있도록 총 542개 문장을 데이터 셋으로 구축하였다.

간호 진술문의 사건 범주 분류를 위해서는 Naive Bayes 분류기를 학습시켰는데, 이 분류기는 진술문이 포함하는 형태소가 태깅된 개별 단어들과 5개의 자질 값들을 이용한다. 이 다섯 개의 자질들의 목록은 아래와 같다.

- is_high_temp
- is_subjective
- is_drug_related
- is_negation
- is_reactive

각 자질의 값은 참 혹은 거짓이다. 'is_high_temp' 자질은 정규식(regular expression)을 이용해 생체정보 기록 진술문에 포함된 체온이 37.4도 이상이면 True를 아니면 False를 갖게 된다. 'is_subjective' 함수는 주관적 진술임을 의미하는 단어, '주관'이라는 단어가 진술문에 포함되었다면 True를 아니면 False를 리턴한다. 그리고 'is_drug_related' 함수는 약물의 처방을 뜻하는 단어 '처방', '투약'이 나오거나 다음의 약물의 이름들이 진술문에 포함되는 경우 True를 리턴하게 된다.

각 약물들의 이름은 ‘keromin’, ‘tyleno’, ‘tramadol’, ‘fentanyl’, ‘brufen’, ‘zithromax’, ‘dexa’, ‘peniramin’이다. 자질 ‘is_negation’은 부정형 표현인 ’없음’을 단순히 포함하면 True 아니면 False를 갖는다. 마지막으로 ‘is_reactive’ 자질은 ADR의 주요 증상인 ‘통증’, ‘구토 혹은 오심’, ‘발열’, ‘어지러움’을 가리키는 표현들이 나타나는 경우 True 값을 갖는 자질이다.

위 다섯개의 자질 셋과 진술문이 포함하는 유니그램들을 자질로 갖는 Naive Bayes 분류기를 학습시켜 총 542개 진술문 데이터 셋을 8대 2비율로 각각 학습 셋과 테스트 셋으로 나누어 분류기의 성능을 테스트하였다. 한 진술문을 7개의 범주 중 하나로 분류하는 이 과제의 최종 결과 정확도는 **0.72** 였다.

7.3.2 보상

보상 함수는 해당 환자가 ADR이 발생한 환자인 경우에는 마지막 진술문에 도달했을 때, $+1$을 부여하게 했고 ADR이 발생하지 않은 환자라면 마지막 진술문 도달시 -1을 주도록 하였다. 그리고 주어진 진술문 상태 기호가 0이나 1이 아니라면 보상값 $+1$을 주도록 했으며 그렇지 않은 모든 경우엔 0을 주도록 했다. 진술문 상태 1과 상태 0을 같이 취급한 이유는 약물 처방 관련 사태 자체는 약물 부작용 여부와 직접적인 관련성이 약한 동시에 약물 부작용이란 사건의 인과적 원인이기 때문이다.

7.3.3 절차

전체 진술문 데이터를 8:1:1의 비율로 나누어 각각 학습, 개발, 테스트 셋으로 나누어, 학습 데이터에 속하는 진술문들을 상태 공간으로 사상하고 이것을 TD(λ) 모형에 따라 상태값을 계산하였다. 이 실험에서 적격 흔적도 문제는 고려 대상이 아니었으므로 replacing 흔적도를 이용하였고, 학습률과 보상 감쇄율(γ)은 각각 0.1로 하였고 흔적 감쇄율(λ)은 0.3으로 모두 낮게 유지하였다.

7.3.4 비교 조건

본 실험은 진술문의 온라인 상태 값을 예측하는 것이 목적이므로 직접적인 분류

Min DF	Max DF Proportion	Gamma	C
5	0.5	0.1	4

표 7.4: SVM 학습에 사용된 최적 모수 값.(C는 regularization 관련 모수, Gamma는 decision boundary 민감도 결정 모수)

과제를 강화 학습에 의해서 수행할 수는 없으나, 그 최종적인 결과를 이용해 분류 과제에 이용하는 것은 가능하다. 이를 위해 개발 데이터 셋을 위에서 서술한 방법으로 사상한 후, 모든 상태에 대한 상태값들의 평균을 환자별로 계산하였다(단, 상태 기호 0의 값은 모두 0으로 계산하였다). 다음으로 간단한 로지스틱 회귀 함수를 학습시키고 이 함수를 이용해 테스트 데이터 셋에 대한 분류(ADR 혹은 non-ADR)를 수행하였다. 이와 비교할 다른 방법들은 근래 분류 과제에 있어 흔히 쓰이는 기계 학습 방밥들로 그 목록은 아래와 같다.

Naive Bayes (NB) 빈도 20 이상의 문장 약 9천개 문장을 사용해서 분류를 위한 입력 자질로 사용하였다. Information Gain을 이용해 입력 자질을 정렬하고 그리드 탐색(grid search) 방법을 통해 최적 수의 자질(N-best features)을 결정하였다. 전체 데이터를 8대 2의 비율로 학습 셋과 테스트 셋으로 나누고, 테스트 셋을 다시 반으로 나누어 검증 셋을 구축하였다. 검증 셋에 대한 그리드 탐색 결과 선택한 3700개의 문장을 사용하였다.

Support Vector Machine (SVM) 역시 빈도 20 이상의 문장을 자질로 하여, SVM에 선형 함수를 사용한 SVM-linear 모형과 Radial Basis Function을 사용한 SVM-RBF 모형을 이용하였다. 각 환자의 진술문 데이터는 TF-IDF 벡터화를 거친 후에 SVM 분류기의 입력 자질 벡터로 활용되었다. SVM 모형의 네 개의 모수에 대해 그리드 탐색을 수행하여 최적 모수 값을 찾은 후에 테스트 데이터 셋에 대한 분류의 정확도를 구하였다. 표 7.4는 이 모수 값들에 대한 정보를 제공한다.

Convolutional Neural Network (CNN) Kim (2014)의 TextCNN 모형을 따라 진술문 데이터를 분류하는 조건이다. 이 모형에서 각 문장들은 일정 길이의 벡터로서 처리되며 여러 층위의 신경망을 거쳐 최종 아웃풋 노드에서 분류가 이루어진다. 그림 7.1은 이 모형의 구조를 시각적으로 알기 쉽게 보여준다. 이 조건에서는 모든 문장(837,293개)을 사용하였고, 문장유형들을 미리 Paragraph

Parameters	Value
Embedding dimension	200
Filter sizes	3, 4, 5
No. Filters	128
Dropout Probability	0.5
Batch size	64
No. Epochs	20

표 7.5: TextCNN 학습에 사용된 모수 값

Vector 모형(Le & Mikolov, 2014)을 통해 pre-trained 한 경우와 그렇지 않은 경우를 모두 실험하였다. 이 실험 조건에서 사용된 모수값들은 표 7.5에 제시되어 있다.

본 실험의 데이터 특성상 장기입원 환자의 데이터는 매우 많은 진술문들로 이루어져 있으므로 주어진 계산비용의 한계상 입력 데이터의 길이를 적당한 크기로 제한하는 작업이 요구되었다. ADR 환자 데이터의 경우 ADR이 발생한 것으로 보고된 날을 이전으로 288개 진술문 리스트를 취하여 이를 입력 벡터로 삼았다. 288이란 모수 값은 ADR이 발생하기까지 생성된 간호 진술문의 평균 갯수에서 비롯되었다. ADR이 발생하지 않은 환자의 데이터 역시 적당한 길이로 편취되어야 했으므로 입원시부터 퇴원시까지의 진술문 리스트에서 무작위로 특정 시점을 선택하여 그 이전의 288개 진술문이 선택되었다. 이 기준 시점이 무작위로 선택된 것은 ADR이 발생하는 시기를 특정할 수 없다는 논리적 가정에 따른 것이다.

Long Short-Term Memory Network (LSTM) 이 조건은 재귀 신경망 모형을 이용하여 진술문 연쇄열에 대해 분류를 시도하는 조건이다. TextCNN 조

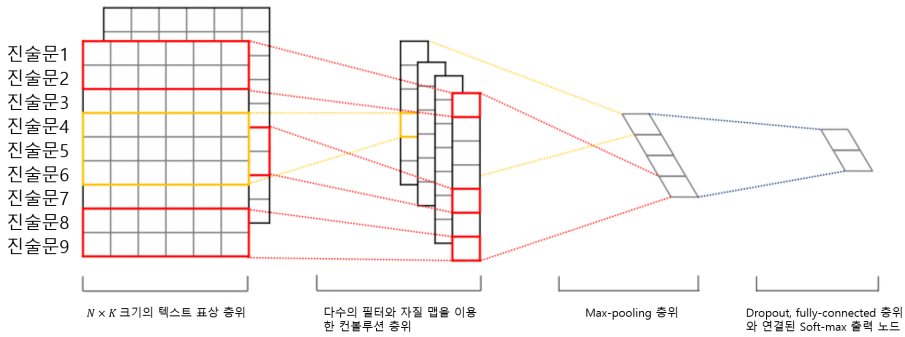


그림 7.1: TextCNN 모형

건에서와 마찬가지로 모든 진술문 유형이 사용되었으며, pre-trained embedding 조건 또한 같이 실험되었다. 또한 한 입력 데이터의 최대 시퀀스 길이가 결정되어야 했으므로 ADR 그룹에 속하는 데이터는 ADR 보고일 이전 200개 진술문 리스트가 입력 데이터로 사용되었고, non-ADR 그룹의 데이터는 무작위 시점의 진술문을 기준으로 마찬가지로 200개 진술문 데이터가 선택되었다. 이 조건의 실험에 사용된 자세한 모수 값 목록은 표 7.6에 제시된 바와 같다.

7.4 실험 결과

표 7.7은 TD 방법을 따라 'unknown' 상태를 제외한 진술문의 각 상태들에 대해 계산된 상태 값들을 보여준다. 이 상태 값의 높고 낮음은 그 상태가 다른 사건의 상태들(상태 2∼6)과 얼마나 가깝게 엮여 나타났는가에 의해 결정된다. 표 7.7에서 상태 1은 다른 진술문 상태들에 비해 현저하게 그 값이 낮았으며, 반면 상태 6은 매우 큰 상태 값을 기록했는데 이는 곧 상태 6이 출현 했을 경우 다른 ADR 관련 상태들이 출현할 가능성이 높았지만 상태 1의 출현 여부는 다른 ADR 관련 사건 진술문들의 출현에 별다른 영향을 미치지 않았음을 뜻한다. 상태 6 다음으로 높은 값을 가진 상태는 2와 4로, 이 상태들은 모두 어떤 고통이나 통증을 호소하는 경우와 밀접한 관련을 보이는 상태라는 공통점이 있다. 상태 3은 의사에게 연락하는 등의 어떤 알림이 일어났을 경우를 주로 가리키는 상태로 이 상태의 값은 다른 상태의 값들에 비해 높지 않았다. 마지막으로 상태 5는 상태 1을 제외하면 가장 낮은 값을 기록 했는데, 이는 이 사건이 의미적으로 약물 유해반응과 연결 가능성이 매우 높다는 점에서 다소 비직관적인 결과이다.

Parameters	Value
LSTM cell size	256
No. LSTM layers	2
Embedding size	200
Dropout Probability	0.5
Batch size	64
Learning rate	0.01
No. Epochs	7

표 7.6: LSTM 분류기 학습에 사용된 모수 값

State	Meaning	Value
0	Unknown event	NA
1	Drug related event	0.17
2	Abnormal reactions	0.47
3	Doctor related event	0.27
4	Subjective response	0.51
5	1 & 2 event	0.18
6	4 & 2 event	0.80

표 7.7: TD 기반 간호 진술문의 상태 값

상태 측정치들의 시간적 추이 변화를 자세히 보기 위해 학습 셋을 다시 구성하여 6,652명의 환자 데이터에 대해 상태 값 측정을 시행하면서 한 명의 데이터 처리가 끝날 때마다 상태 값들을 기록하였다. 그림 7.2는 이 값들의 그래프로 X축은 랜덤하게 섞인 환자별 데이터의 순차 인덱스를 가리킨다. 이 그래프에 따르면, 학습 중반부 이후 상태 6의 상태 값의 안정화 현상이 두드러지게 나타난다. 이는 표 7.7에서 관찰된 상태 6에 대한 측정치가 다른 상태 값들에 비해 신뢰할만한 것임을 보여주며, 상태 6이 출현할 경우 다른 ADR 유관 진술문들이 높은 확률로 뒤따라 나타났음을 뜻한다. 반면에 다른 상태들은 상태 6에 비해 복잡한 동역학적 관계를 보여주었으나, 각 상태들의 상승과 하강은 비슷한 시간 윈도우 상에서 발생한다는 것을 관찰할 수 있었다. 즉, 상태 6을 제외한 다른 상태들은 서로 깊은 시간적 의존성을 가지는 것으로 보인다. 따라서 이들 상태들은 서로 짧은 시간차를 두고 출현하지 않으면 서로 상태 값이 동반 하락하는 패턴을 갖는다.

표 7.7의 상태 값에 근거하여 검증 데이터셋(validation set)에서 ADR 조건과 non-ADR 조건 별로 등장한 진술문들의 상태 값 평균을 계산한 결과, ADR 조건은 0.12를 기록하였고 non-ADR 조건은 0.1을 기록하여, ADR 조건이 non-ADR 조건에 비해 평균 상태 값이 높게 나타났다. 즉, 이는 ADR 유관 문장이 ADR 환자 조건에서 더 자주 출현했음을 뜻한다. 본 실험에서는 이 평균 값을 분류 자질로 활용하여 로지스틱 회귀를 통해 ADR 예측 과제를 수행할 수 있는지를 알아보았다. 그리고 그 결과는 앞에서 기술한 다른 비교 방법들의 결과와 함께 표 7.8에 제시되었다. 가장 높은 정확도를 기록한 것은 Naive Bayes 모형이었으나 SVM 결과와 의미 있는 차이를 보이지는 않았다. TD 기반 로지스틱 분류 모형은 매우 단순한 알고리즘을 사용하였음에도 CNN 모형의 정확도 보다 높았으며 LSTM

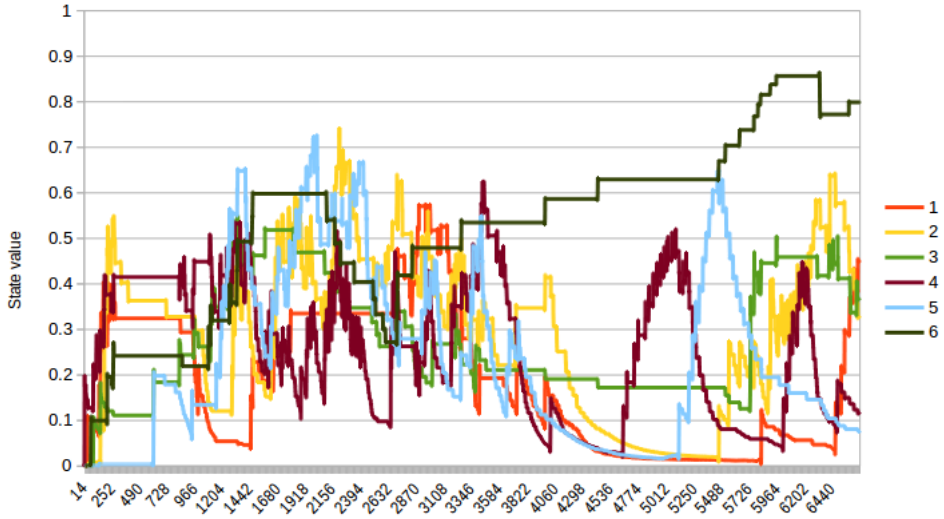


그림 7.2: 학습 데이터에 대한 상태 값의 추이 그래프

Method	Accuracy
NB	0.64
SVM (linear)	0.63
SVM (RBF)	0.63
CNN	0.58
CNN (with embedding)	0.58
LSTM	0.61
LSTM (with embedding)	0.57
TD-based logistic regression	0.61

표 7.8: ADR 분류 과제 정확도 실험 결과

모형의 정확도와 같은 정확도인 0.61을 기록하였다.

제 III 편

논의 및 결론

제 8 장

단어 감정 극성값 측정법 논의

시간 차 학습을 이용한 단어 감정 극성값 측정 실험들을 통해, 이 방법이 의미 있는 측정치를 제공한다는 것을 앞에서 살펴 보았다. 본 장에서는 이 결과들의 의미를 보다 자세하게 분석하고 논의하고자 한다.

8.1 TD 방법들 간의 분산의 차이

True Online TD를 제외하면 전반적으로 TD 방법들의 수행이 비슷한 수준을 기록하였으나, 학습률의 변화에 따른 수행도에서는 뚜렷한 차이점이 나타났다. 테이블 8.1은 각 TD 방법들이 학습률에 대해 보였던 결과들의 분산을 보여주는데 significance 흔적도를 이용한 조건이 가장 낮은 분산을 기록한 것을 확인할 수 있다. 이 점은 그림 6.3에서도 시각적으로 확인되는데 다른 조건들에 비해 TD(1)의 수행 수준이 학습률의 증가에도 하락하지 않는 듯한 패턴이 나타난다.

이 현상은 significance 흔적도 조건이 발생 빈도가 매우 높은 단어들에 대해 매우 낮은 흔적도를 부여한다는 점에서 어느 정도 설명이 가능해 보인다. 시간

Method	Average of VARs
TD(λ) with accumulation	0.0006
TD(λ) with replacing	0.0002
TD(λ) with Dutch	0.0002
TD(λ) with significance	**0.0001**
True Online TD(λ)	0.0002

표 8.1: 학습률(0.1∼0.5)에 대한 TD(λ) 성능의 분산 평균값

차 학습에서 상태 값의 갱신은 시간 차 오차(TD-error)에 학습률과 흔적도를 곱한 값에 의해 이루어지는데($V(s) = V(s) + \alpha\delta e(s)$), 학습률이 크고 또한 특정 시점에서의 상태 흔적도 값이 크다면 한 번에 매우 큰 상태 값 갱신을 불러올 수 있기 때문이다. 그리고 이것은 결과적으로 학습과정에서 학습률의 증가에 민감한 요소로 작용할 여지가 충분하다.

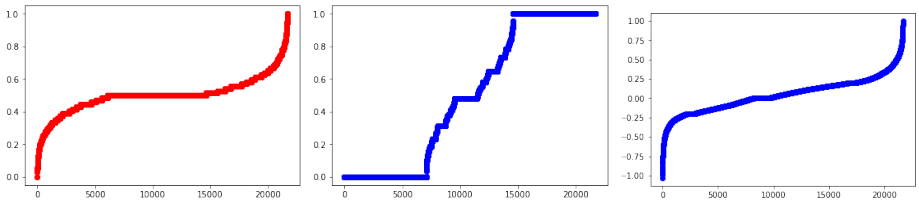
또한 이는 accumulation 흔적도와 Dutch 흔적도 조건이 왜 학습률의 상승에 따라 급격한 성능 하락을 보였는지를 부분적으로 설명한다. 이들 흔적도 체계에서 고빈도의 단어들의 흔적도는 누적적으로 값이 크게 증가할 수 있기 때문에, 학습률 마저 큰 경우 상태 값 갱신의 크기 정도는 이에 비례하여 커지게 되고, 따라서 이는 상태 값의 갱신 및 수렴 안정성을 저해할 수 있을 것이다.

그러므로 자연어 텍스트의 단어에 상태 값의 측정 방법으로써 TD 방법을 적용한다면 replacing 흔적도나 significance 흔적도를 사용하는 것이 유리할 것이다. 만일 학습률을 결정하기 힘들다면 significance가 안전한 대안이 될 것이고, 계산적으로 경제적인 선택지를 선호 한다면 replacing 흔적도가 더 나은 선택이 될 것이다.

8.2 TD 기반 측정치의 평탄성

섹션 6.3의 실험에서 TD 기반 측정치와 단어의 감정 극성 값과 직접적인 비교 실험 결과를 보고 하였다. 그리고 그 결과, 베이즈 확률 에 비해 TD 기반 측정 값이 보다 높은 상관관계를 보였는데, 그림 8.1는 그 배경적 요인을 잘 보여준다. 이 그래프는 측정된 값을 오름차순으로 그래프로 표현한 것으로, TD 방법의 측정치가 비교 방법에 비해 그 패턴이 보다 평탄하다는 (smoothing) 것을 잘 나타낸다. 그리고 이 패턴은 흥미롭게도 주석된 감정 극성 값의 그래프와 유사하게 나타난다.

베이즈 확률은 빈도가 적은 어휘에 대해서는 0 아니면 1 식의 극단적 확률을 제공하는 것으로 보이며, 확률의 증가 양상도 TD 방법에 비하면 상대적으로 평탄하지 않은 것처럼 보인다. 이는 그래프 패턴의 선형적 증가 양상을 저해할 것이므로 Pearson 상관 계수를 낮추는 하나의 요인으로 작용할 것으로 여겨진다. 따라서 TD 방법에 의한 측정이 측정 값들끼리 서로 그 크기가 비교 가능한 결과를



(a) Annotated (b) Bayesian (c) TD-based

그림 8.1: Stanford Sentiment Treebank 데이터의 전체 어휘 셋에 대해 주석된 극성 값(annotated), 베이즈 확률 값(Bayesian), 그리고 TD(1) 기반 극성 값(significance traces)의 데이터 플롯

제공한다는 점에서 본 실험의 베이즈 확률에 비해 이점을 가진다.

8.3 TD 방법의 점진성

시간 차 학습의 참된 강점은 그러나 단순히 어떤 극성 값을 정확하게 측정할 수 있다는 데 있는 것이 아니라, 그 측정이 시간의 흐름에 따라 점진적(incremental)으로 이루어질 수 있다는 점에 있다. 즉, 이 방법은 순전하게 실시간으로 어떤 어휘의 감정 극성 값을 모니터링할 수 있도록 돕는다. 그림 8.2은 그 이점을 잘 보여주는 실용적 예제이다.

이 예제는 영화 리뷰 사이트(`https://www.rottentomatoes.com`)에서 수집한 400개 영화평에 TD(1) 방법(siginificance traces)을 이용하여 특정 단어('spielberg')의 감정 극성 값을 매 영화평 마다 기록한 것이다. 이 400개의 영화평은 스필버그가 감독한 각 4개의 영화에 대해 100개씩 수집 되었다(Jurassic Park(JP), The Lost World: Jurassic Park (JP2), Saving Private Ryan (SPR) and A.I. Artificial Intelligence (AI)).

이 예제에서 볼 수 있듯이 스필버그의 가장 성공적인 두 영화(Jurassic Park, Saving Private Ryan)에서 가장 큰 상태 값의 피크(peak)가 0.8에 근접하여 발생하는 것을 볼 수 있다. 어떤 인물 혹은 대상의 인기도란 사실 시간적, 추세적 성격을 갖는 것이므로 TD 방법을 통한 극성 값 측정은 이런 점을 손쉽게 포착해준다는 점에서 전통적인 통계적 방법보다 유리하다고 볼 수 있는데, 이는 시간 차 방법이 본질적으로 점진적 성격을 갖는 알고리즘이기 때문이다.

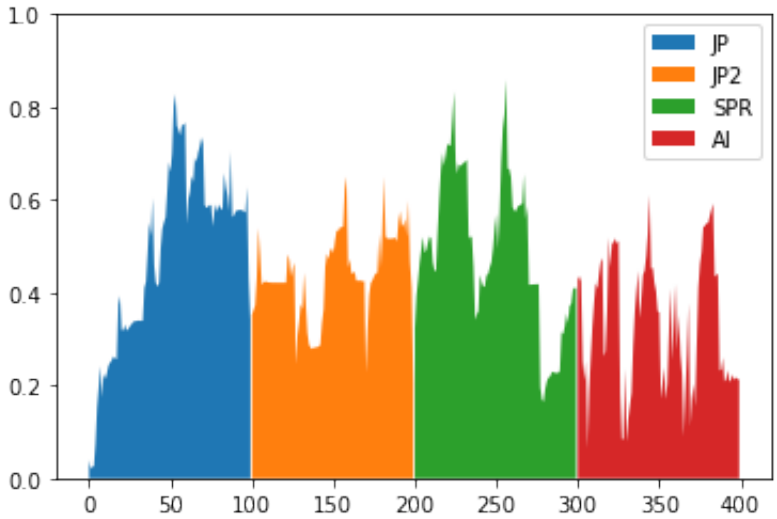


그림 8.2: 4개의 스필버그 영화에 대한 400개 영화평으로부터 단어 'spielberg'의 감정 극성 값 추적 그래프

제 9 장

간호 진술문 상태 값 측정법 논의

본 논문에서 간호 진술문의 상태 값 측정법은 텍스트의 온라인 상태값 측정 과제의 맥락에서 연구되었다. 앞의 6장에서 실험되고 8장에서 논의된 감정 극성 값 측정 과제에서 측정 대상은 단어였기 때문에 함수 근사법(function approximation)이 사용될 필요는 없었다. 하지만 간호 진술문 상태 값 측정 문제에서 측정 대상은 문장이기 때문에 함수 근사법 사용이 고려되어야만 했는데, 이는 문장이 생성될 수 있는 경우의 수가 거의 무한하기 때문이다. 이로 인해 말미암는 문제와 희박한 보상 값의 문제를 해결하고자 기억 기반 함수 접근법과 더불어 모형 기반 강화학습이 이용되엇다. 기억 기반 접근은 ADR과 이미 관련 있는 것으로 알려진 진술문에 초점을 맞추어 ADR과 상관이 약한 진술문들에 대한 탐색의 부담을 크게 줄여준다. 또한 본 연구에서는 희박한 보상 값의 문제를 우회하고자 보상함수을 모형 기반적 방식으로 설계하였는데, 이는 ADR과 관련된 관찰자의 지식을 강화학습의 모형에 반영한 것이다.

9.1 문장에 대한 함수 근사법의 문제

문장보다는 더 단순한 단위인 단어의 임베딩 모형은 자연어 처리의 역사에서 다양한 방법으로 연구되고 발전되어 왔는데, 전통적으로는 LSI (Latent Semantic Indexing), Word-Space model이 대표적이다 (Schütze, 1998, 1993; Salton, Wong, & Yang, 1975; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). 이에 비해 문장을 벡터 공간으로 사상하고자 하는 문장 임베딩에 대한

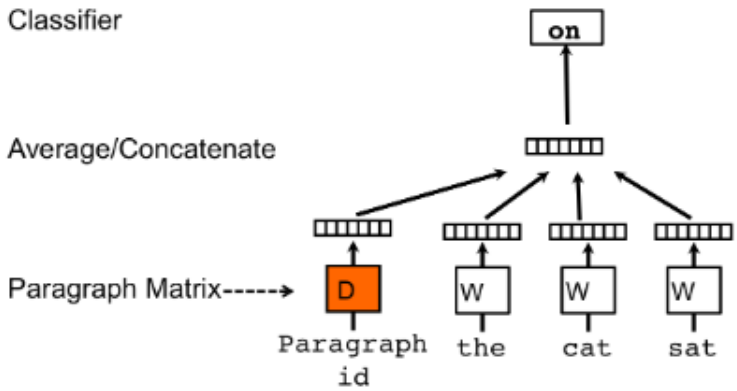


그림 9.1: Distributed Memory of Paragraph Vector (DM-PV 모형)

연구는 상당히 최근에야 이루어졌다. Le & Mikolov (2014)이 제시한 Paragraph Vector Model은 기존의 단어 임베딩 모형 (Mikolov, Chen, Corrado, & Dean, 2013)을 문장 혹은 문서에 대해 확장해서 적용하는데, 문서 벡터를 생성하여 단어 벡터들과 함께 학습시켜 다음 단어를 예측하는데 있어 문서 벡터와 앞선 단어 벡터들의 맥락이 얼마나 도움이 되는지 정도를 반영하여 문서 벡터를 학습시킨다 (그림 9.1).

그리고 위와 유사하지만 인공 신경망 모형을 이용한 대표적인 문장 임베딩 연구가 있다. Kiros 외 (2015)의 Skip-thought Vector 모형은 단어 임베딩 모형과 유사하게 문장 앞 뒤 맥락을 이용하여 문장을 벡터 공간으로 표상시키도록 고안되었고, 이들 모형에선 RNN 기반 encoder-decoder 방법을 이용해 문장을 인코딩한다.

위 문장 임베딩 방법들은 공통적으로 문서의 감정 분류에 있어서 문서 벡터 정보가 도움이 되었는지를 봄으로써 모형의 타당성을 검증한다. 하지만 강화 학습의 함수 근사법이 필요로 하는 종류의 상태 표상을 문장 임베딩이 제공하는지는 아직 알려지지 않았다. 따라서 본 연구는 보다 계산적으로 간단한 모형인 DM-PV 모형을 이용해 문서 검색(document retrieval) 과제를 수행하였다. 표 9.1의 결과는 아주대학교 간호 진술문 데이터로부터 고빈도 문장들을 추출하여 비슷한 의미를 가진 문장들끼리 묶어 분류한 데이터로부터 얻은 것이다. 간호사 경력이 있는 간호 진술문 전문가가 수작업으로 총 14,260개 문장들에 그 의미 유형을

	Paragraph Vector
P@10	0.28
P@15	0.27
P@20	0.27
MAP	0.27

표 9.1: DM-PV 모형을 이용한 문서 검색 수행도

주석하였고, 그 중에서 최소 20개의 동일한 의미 유형을 가진 114개 의미 그룹을 추출했다. 그리고 각 의미 그룹으로부터 한 문장을 랜덤으로 추출하여 총 114개 질의용 문장 셋을 구성하였다. 문서 검색 과제는 각 질의 문장으로부터 얼마나 많은 동일한 의미 그룹의 문장들을 검색하였는지를 정밀도(precision)로 판단하여 시행되었다.

표 9.1는 이 결과를 보여주는데, $P@K$는 K개 정보 검색에서 쿼리와 유관한(relevant) 항목을 추출한 비율을 가리키며 MAP는 Mean Average Precision을 뜻한다. 이 결과에 따르면, 정밀도는 0.27 수준으로 높지 않은 것으로 나타났으며, 이는 문장 임베딩을 통해 함수 근사법을 시도하기에는 위험성이 크다는 것을 뜻한다. 본 연구는 이러한 복잡성 문제를 회피하고자 기억 기반 접근법(memory-based approach)를 택했으며, 앞의 7.3.1에서 기술된 바 이 방법의 정확도는 0.72를 기록하였다.

연구 배경(제 5장)에서 기술했듯이, 강화학습에서 기억 기반 접근법은 미리 준비한 학습 자료로부터 주어진 상태에 대한 상태 값을 이끌어내는 방법을 가리킨다. 본 연구에서는 주어진 진술문의 상태(query state)의 상태 값을 직접 이끌어내는 방식으로는 이 접근법을 활용하지 않았으며, 다만 주어진 진술문의 사건적 유형(event class)를 알아내는 목적으로만 이용하였다. 그 주된 이유는 어떤 사건 유형이 어떤 상태 값을 갖는지에 대한 모형이 부재하기 때문이며, 다른 부차적 이유는 진술문이 가리키는 사건에 대한 가치 예측은 모형 자유(model-free) 방식으로 이루어지기를 의도했기 때문이다. 그러나 이 문제의 환경상 주어지는 보상 값이 매우 희소했기 때문에 보상 값 부여 체계에 모형을 도입해야 했다. 본 연구는 ADR 유관 문장들이 짧은 시간 차이로 인과적으로 연결될 경우, 더 큰 보상 값이 각 유관 상태들에 주어지도록 강화학습 보상 함수를 설계하였다. 그와 함께

자연스럽게 본 연구의 방법은 상태들의 이행양상에 대해서는 모형 자유적 접근을 취하고, 보상 값 부여에 있어서는 모형 기반적 접근을 취하는 유사-모형 기반적 강화학습의 형태를 갖게 되었다.

9.2 진술문 상태 값의 점진적 측정

본 논문에서 제시한 방법의 간과되지 말아야 할 중요한 장점은 이 방법의 점진성(incrementality)에 있다. 즉, 본 연구의 방법은 간호 진술문이 입력되는 매 순간마다 약물 부작용 위험도가 얼마나 높은지에 대한 정보를 제공할 수 있다. 또한 본 연구의 설계에서 보상 체계(reward system)는 모형 기반이지만, 상태 전이는 모형에 의존하지 않으므로(model-free) 실제 학습 환경이 어느 정도 변화하더라도 진술문에 대한 상태 값 추정 역시 그에 따라 변화 가능하다는 중요한 장점을 갖는다.

표 9.2은 이러한 점을 쉽게 관찰하기 위한 예제로 실제 진술문 문장과 그에 맞게 변환된 상태 기호들을 보여준다. 그리고 그림 9.2은 앞의 실험으로부터 얻은 상태 값 벡터를 입력으로 하여 TD(λ) 알고리즘으로 계산한 상태 값의 시간적 추이에 따른 상태 값의 그래프이다. 그림에서 볼 수 있듯이, 0이 아닌 진술문 상태 사이에 0 기호를 가진 상태가 등장하더라도 값이 하락이 크지 않은 것은 다음 상태 값의 부트스트래핑에 따른 효과이다. 시간차 학습 알고리즘의 이와 같은 특성은 진술문 상태 값의 예측에 있어 이해하기 쉬운 직관적 도구를 제공한다.

만일 ADR 유관 상태들이 그림 9.2에서와 달리 상태 0을 사이에 두고 서로 떨어져 있다면 시간차 학습 특성상 ADR 유관 상태들은 평균적으로 낮은 상태 값을 갖게 된다. 따라서 본 과제에서 시간차 학습은 ADR 발생 유무에 대한 추론 양상을 직관적/인지적 방식과 유사하게 모델링하는 셈이 된다.

9.3 진술문 상태들의 관계에 대한 TD 방법의 유연성

본 연구에서 TD 방법의 추가적 장점 하나는, 진술문 상태들이 서로에 대해 갖는 인과적 관계의 거리를 사전에 알 경우 이를 보상 감쇄율(γ)를 통해 조절할 수 있다

Sentence	State
침습적시술	0
감염의 증상이 없다부종,발적,동통 증상없음	0
두통 감소함	0
정맥주사부위 통증,부종,발적 없음	0
침습적시술	0
감염의 증상이 없다	0
감염 증상 없음	0
관 삽입부위 분비물 없음	0
활력징후 측정함BT: 37.2도	0
주관적 진술:영양제 바꾸고 근육통이랑 근육이 뻣뻣해졌어요	6
수액 주입중임내용:Oliclinomel+MVH	0
사지 뻣뻣해지고, 근육통 호소함	2
의사에게 알림의사:XXX	3
의사 XXX fluid stop하고 지켜보자함	1
통증 없음	0
삽입관 및 부착물 있음부위:abdomen, 상태:음압, 종류:baro-vac	0
금식 중임	0

표 9.2: 간호 진술문 문장과 그 상태 기호 테이블

는 사실이다. 관찰하고자 하는 현상의 성격에 따라 상태들이 갖는 인과적 거리는 다를 수 있는데, 예들 들어, 병원에서 받는 물리적 수술 같은 현상이 수술 부작용 사건에 대해 갖는 인과적 영량력의 거리는 매우 길 수 있으나, 식사와 같은 사건이 소화불량과 같은 사건에 대해 갖는 인과적 거리는 그보다 훨씬 짧을 것이다.

TD 방법의 보상 감쇄율(γ)은 이를 조절하는데 편리한 방법을 제공한다. 만일 사전에 상태들이 갖는 인과적 영향력의 거리가 짧다면, 감쇄율 값을 0에서 1 사이에서 낮게 유지하면 되기 때문이다. 반면에 그 인과적 영향력의 파장이 길다고 가정하면 1에 가깝게 그 모수 값을 정하는 식으로, 관찰하고자 하는 현상의 성격에 따라 조절할 수 있다.

표 9.3는 본 연구의 학습 데이터의 500개 케이스로부터 학습시킨 결과로서, 약물 부작용 현상에 있어 보상 감쇄율의 변화가 미치는 영향을 보여준다. 그리고 실제로 약물 부작용 사건이 갖는 인과적 파급력의 거리는 짧기 때문에 감쇄율을 0.1로 정한 경우에 훨씬 분별 가능한 값들이 나온 것을 관찰할 수 있었다. 이와 반대로 감쇄율을 1에 가깝게 하면 상태들의 값들은 거의 비슷해지는 것을 볼 수 있는데, 이는 약물 부작용과 연관된 사건들은 장기적 맥락에서는 서로 인과적

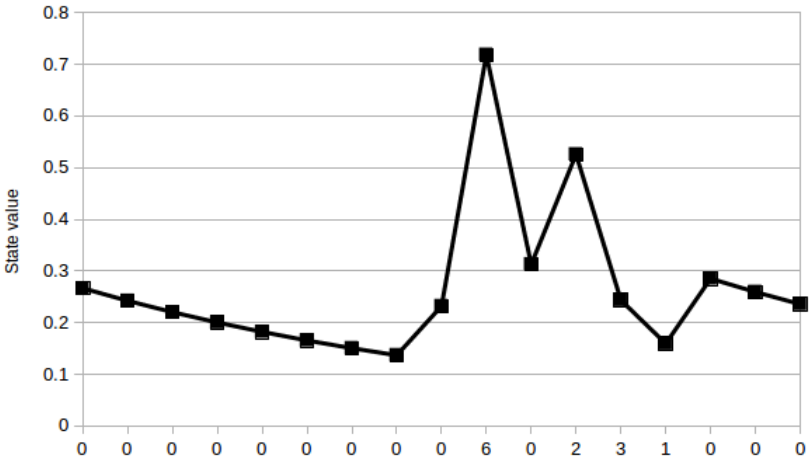


그림 9.2: 진술문의 시간적 연쇄에 따른 상태 값 그래프

State	1	2	3	4	5	6
γ=0.1	0.045	0.51	0.19	0.08	0.39	0.78
γ=0.99	20.95	22.17	21.64	20.44	20.94	22.6

표 9.3: 보상 감쇄율에 따른 상태 값 변화 테이블

상관이 거의 없다는 것을 의미한다.

9.4 한계점 및 도메인 특정적 장점들

기억 기반 함수 근사법 및 모형 기반적 강화학습 방법은 분명 그 장점이 있으나, 반대급부도 여럿 존재한다. 우선 기억 기반 접근법의 주요 문제는 학습을 위한 예제 자료들을 확보하는 것과 그 학습자료가 신뢰받기 어려울 경우, 상태 표상 및 상태 값 예측에 수정 불가능한 오류가 생기게 된다는 점이 있다. 본 연구에서는 이 문제를 가능한 최소화 하기 위해 랜덤 샘플링을 통해 학습 예제들을 추출하고, 사건 공간의 크기(7개)를 최대한 낮게 유지하고자 하였다. 하지만 그럼에도 진술문의 상태 공간의 사상의 정확도는 완벽하기를 기대하기 어렵다. 따라서 이 과정의 부정확성에서 비롯된 문제점은 여전히 남아 있다.

모형 기반 강화학습의 문제 역시 이와 유사하게 '모형의 정확성'에 있다. 만일 모형이 정확하지 못하면(본 연구의 경우 보상 값 제시가 부정확하다면), 상태 값에 대한 정확한 예측을 하지 못하므로 강화학습은 문제 해결에 대한 효과적인

해결책을 제시하지 못하게 된다. 본 연구에서는 이 문제에서 발생하는 오류는 진술문의 상태 공간에로의 사상 과정에 많이 편중되어 있는데, 왜냐하면 보상 값 제시 방식이 세상사 지식(world knowledge)에 의한 결정적(deterministic) 규칙을 따르므로 가장 큰 부정확성은 진술문이 잘못된 사건 상태에 속하게 될 때 발생하게 된다.

하지만 실용적 측면에서, 그리고 의료라고 하는 도메인의 특별한 환경에서 오는 장점은, 많은 병원에서 간호 진술문은 전문적 관리의 대상이라는 점이다. 전문 인력에 의해 간호 진술문은 그 유형이 정제되거나 특정한 수로 제한될 수 있고, 각 진술문이 함의하는 사건 정보도 메타 태그 형태로 관리될 수 있다. 이런 다소 이상적인 의료 정보 시스템 내에서 본 연구의 방법론은 진술문의 상태 표상 과정에서 오는 오류를 거의 제거할 수 있다. 또한 약물 부작용과 같은 사건의 발생 시점과 같은 정보도 의료 정보 시스템 내에서 정확하게 입력될 수 있으므로, 강화 학습의 보상 함수 관련한 기제도 더 정확해질 수 있다. 또한 본 연구의 방법론은 이 방법의 적용 대상이 반드시 약물 부작용일 것을 요구하지 않는다. 같은 입력 정보에 대해서 보상 함수와 목표하는 사건 상태 공간을 달리 설정하는 방식으로, 다른 많은 유사한 문제들에도 본 방법을 적용할 수 있다. 그리고 이런 점은 지도 학습방식처럼 목표로 하는 대상마다 학습 데이터와 테스트 데이터를 준비하여 대응해야 하는 번거로움에 비하면 상당한 장점이 될 수 있다.

제 10 장

전체 논의 및 결론

본 논문은 텍스트의 온라인 처리라는 측면에서 강화학습을 응용하는 방법을 실험하고 논의하였다. 먼저 단어 감정가의 온라인 측정법에 시간차 학습 알고리즘인 TD(λ)를 적용하는 방법을 연구하였고, 이어서 같은 알고리즘으로 약물 유해반응과 연관된 간호 진술문의 상태 값을 측정하는 방법을 살펴 보았다. 이 과정을 통해 강화학습을 자연어 처리에 이용하는데 있어 몇 가지 난점들이 드러났다. 본 연구에서 언급된 가장 큰 난점은 바로 **보상 값의 희소성**으로, 이 문제를 해결하는 것은 강화학습 기반 자연어 처리의 성공과 밀접한 연관을 갖는 것으로 보인다. 본 장에서는 이 문제를 중심으로 앞서 논의된 연구들을 요약하고, 그 의미를 정리하고자 한다.

10.1 희소한 보상 값의 문제

Chomsky(1959)가 일찌기 지적했듯이, 사람들은 언어 학습시 언어 사용과 관련된 특별한 보상 혹은 처벌을 행동주의적 방식으로 받지 않는 것처럼 보인다. 그리고 우리가 언어를 언어 행위와 그에 따른 보상이나 처벌을 통해 배우지 않는다면 가장 그럴듯한 설명은 언어 학습을 가능하게 해주는 인지구조를 인간이 갖고 태어나기 때문이라는 것이다. 그리고 이 희소한 보상 값의 문제는 강화학습 방법론을 통한 자연어 처리에도 그대로 적용되는 것처럼 보인다. 강화학습의 이론적 프레임워크에서 인지행위자, 즉 에이전트는 자신의 행위를 통해 환경과 끊임없이 피드백을 주고 받는다. 그러나 만일 이 피드백이 부실하다면, 즉 에이전트의 행동이

좋은 것인지 나쁜 것인지를 알려주는 정보가 부족하거나 희소하게 주어진다면, 에이전트는 자신의 행위를 최적으로 조절하기 위한 기반을 확보하지 못하게 되는 것이다.

본 연구의 시간차 기반 단어 감정가 측정 실험을 통해, 보상 값이 희소하게 주어지는 경우(텍스트의 말미에서 $+1$ 혹은 -1)와 단어 이행 때마다 주어지는 경우를 실험하였다. 그리고 결과적으로 실제 보상 값이 희소하게 주어졌을 경우에 시간차 방법은 Information Gain 기반 단어자질 선별방법과 거의 비슷한 성능을 보였지만, 보상 값이 매번 주어졌을 때는 Information Gain을 상회하는 수행을 보였다.

간호 진술문 상태 값 측정 실험에서는 이러한 희소한 보상 값 문제가 더 심각해지는데, 왜냐하면 한 사람에게 주어지는 진술문의 개수가 매우 많고 약물 부작용과 연관된 진술문을 한 두개로 특정하기 어렵기 때문이다. 그런 이유로 인해 본 연구에서는 모형 기반 강화학습 접근법을 이용하여 강화학습 환경의 보상 메커니즘을 설계하여 문제해결을 시도하였다. 여기서 사용된 해결법은 기억 기반적 접근법에 의존하여 현재 상태와 목표 상태와의 유사한 정도를 보상기준으로 활용하는 것이었다. 그리고 이와 같은 해결책은 희소한 보상 값의 문제를 완화하고 진술문의 연속적 상태를 실시간으로 모니터링 하는데 도움을 준다는 것을 실험을 통해 관찰하였다.

하지만 희소한 보상 값의 문제를 모형 기반 강화학습을 도입해서 해결하려는 시도는, "바로 거기에 필요한 그 모형을 어떻게 얻는가?"라는, 또 다른 문제를 남기게 된다.

10.2 모형 기반 강화학습과 자연어 처리의 문제

이 섹션에서는 감정 분류(sentiment classification)라는 자연어 처리에서 접하기 쉬운 과제에 모형 기반 강화학습을 이용하려 할 때, 어떤 문제와 부딪히는 지를 간단한 실험을 통해 언급하고자 한다.

가장 먼저 강화학습을 위해서는 상태가 정의되어야 하는데 Bag of Words 모형을 써서 주어진 텍스트에 대한 바이너리 벡터를 상태로 정의하였다. 다만

Time	Text	Label
1	Effective	positive
2	Effective but	negative
3	Effective but too-tepid	negative
4	Effective but too-tepid biopic	negative

표 10.1: 강화학습 에이전트를 위한 Naive Bayes 모형의 작동 예제

Method	Accuracy
Naive Bayes	0.45
Q(λ)	0.40

표 10.2: 감정 분류 과제 정확도 결과

상태 공간의 크기를 가능한 최소로 하기 위해 나타난 단어 대신 그 단어의 문법적 품사 태그를 그 필드로 삼았다. 하지만 부정어(negation)나, 감정가를 갖는 형용사, 'but'과 같은 맥락 변화에 영향을 주는 경우는 품사 정보를 필드로 삼지 않고 특별히 그에 맞는 필드를 따로 정의하여 사용하였다. 그러므로 전체 상태 공간의 크기는 53개로 크지 않았다.

그리고 전체 문장이 아니라 문장의 일부분에 대해서도 보상 값을 줄 수 있는 사전에 학습된 모형이 필요하므로, Naive Bayes 분류기를 텍스트의 바이그램들을 자질로 삼아 학습시켜 이를 강화학습을 위한 모형으로 이용했다. 표 10.1는 한 짧은 영화평에 대해 이 분류기가 제공하는 지도용 데이터의 예이다. 이 실험의 데이터는 Stanford Sentiment Treebank를 사용했으며, 학습 셋, 검증 셋, 테스트 셋의 구분은 데이터가 제공하는 방식 그대로 재사용하였다(Socher et al., 2013).

주어진 상태에 대해 에이전트에게 주어진 행위공간의 크기는 3(positive, negative, neutral)이었다. 보상 함수 정의는 단순하여 만일 에이전트가 취한 분류행위가 NB 모형의 결과와 갖다면 보상 1을, 다르다면 보상 0을 제공하였다. 이 방법을 통해 과연 강화학습 에이전트는 분류과제를 배울 수 있을까? 본 연구에서는 Q(λ) 알고리즘(Watkins & Dayan, 1992)과 experience replay 방법을 이용하여 매 순회마다 무작위로 샘플링한 20개 데이터를 재학습 시키는 방식으로 학습을 수행시켰다.

표 10.2는 위 탐색적 질문에 대한 결과를 보여준다. 그리고 그 결론은 제한적 긍정에 가까운 것으로 보이는데, 강화학습 에이전트는 감정 분류를 학습하기는

하였으나 그것을 지도한 NB 모형보다 낫거나 같은 성능을 보이지는 못했다. 문제는 부트스트래핑이 특별히 장점으로 작용하지 않는 종류의 과제라면, 강화학습 에이전트가 자신을 지도한 모형에 비해 더 나은 성능을 가질 것이라고 기대하기가 어렵다는 점이다.

또한 강화학습의 중요한 특징인, '시행과 오차'를 통한 학습방식은 지도학습 방법에 비해 더 긴 학습시간을 요구하게 되는데 이는 행위 공간의 크기가 크면 클수록 더욱 길어질 것이다. 이런 문제들은 강화학습을 자연어 처리의 여러 과제에 적용하고자 할 때 흔히 부딪힐 수 있는 문제들로, 이 문제의 극복을 위해서는 보다 효과적인 모형 기반 강화학습이 필요할 것이다.

10.3 본 연구에서 밝힌 것

본 논문의 주요 결론을 요약하면 다음과 같다.

- 강화학습 방법론은 온라인 처리를 지향한다는 점에서 인간의 언어처리 기제와 유사한 측면이 있다.
- 강화학습 방법은 실시간 처리 혹은 실시간 분석을 요하는 종류의 과제에 효과적으로 적용될 수 있다.
- 모형 기반 강화학습은 보상의 희소성 문제를 해결하는데 있어 효과적인 대안이다.
- 모형 기반 강화학습은 인지적 처리 모형을 구현하는데 있어 유용하다.

10.4 강화학습을 이용한 자연어 처리의 미래

본 논문의 서두에서 본 연구의 주요 연구 동기가 강화학습 방법론을 자연어 처리에 어떻게 도입할 것인지에 대한 질문에 답하는 것이라고 밝혔었다. 그리고 여러 실험을 통해 이 과정이 쉽지는 않음을 보였는데, 이 어려움의 근원은 단순 강화학습의 규칙을 적용하여 과제들을 해결하기에는 언어 현상이 복잡하기 때문인 것으로 여겨진다. 이 어려움은 강화학습을 자연어 처리 과제에 이용하고자 하는

근래의 연구들에서 공통적으로 관찰되고 있다. 그리고 이런 시도들에서 보이는 하나의 공통적인 경향은 이 연구들이 인지적 모형을 적극적으로 활용하는 학습 전략을 사용한다는 점이다.

그 중 대화 시스템에서 주로 연구되는 중요하고도 오래된 연구 주제인, 질의 응답(Question & Answering) 관련 연구에서 자연어 질의어를 의미 질의어(logical query)로 파싱하는데 강화 학습을 응용하려는 시도들이 있다(Liang, Berant, Le, Forbus, & Lao, 2017; Zhong, Xiong, & Socher, 2017; Neelakantan, Le, Abadi, McCallum, & Amodei, 2017; Mou, Lu, Li, & Jin, 2017). 이중에서 특히 Liang et al. (2017)은 QA 시스템을 크게 'Manager', 'Neural Programmer' 그리고 'Symbolic Computer' 파트로 나누었는데, Neural Programmer 모듈은 자연어 질의어를 인공 신경망 모형을 통해 논리 구문 형식의 명시적 표상 형식으로 번역하고, 이를 Computer 모듈이 실행하여 그 성공 혹은 실패 여부를 알리고 Manager 모듈은 그 결과에 따라 보상값을 Programmer 모듈에 전달한다. 그럼이 Programmer 모듈은 강화 학습의 REINFORCE 알고리즘(Williams, 1992)에 따라 자신의 질이의 번역 프로세스를 더 많은 보상값을 받을 수 있는 방향으로 시맨틱 파싱을 수행한다. 이 시스템은 질의어의 의미를 상태 표상으로 사상하는 과정에서 재귀 선경망 모형인 Seq2Seq 모형을 통해 상태 표상을 수행하고, 또한 복수의 가능한 번역 형태가 존재할 경우 가장 최선의 선택을 강화 학습 알고리즘에 의존하여 수행한다. 이는 질의어 처리 과정에 대한 인지적 모형화라고 해석할 수 있으며, 상태 표상과 같은 과정에는 지도 학습 기반 알고리즘을 활용하고 보다 이산적이고 명시적인 의사결정 과정에는 강화 학습 알고리즘을 활용했다는 점에서, 인지모형 기반 시스템이라고 볼 여지가 충분하다.

본 논문의 간호 진술문의 상태 측정법 연구에서도 위와 같은 인지적 직관을 모형에 도입하는 것이 상당히 유용함을 보았다. 인지모형을 통한 접근은 상태 공간이 탐색하기에는 너무 방대할 경우에 효과적일 수 있는데, 이는 우리의 인지적 직관이 제공하는 상태 차원 감소(dimensionality reduction)와 보상 체계를 십분 이용할 수 있기 때문으로 보인다. 또한 이러한 인지모형 기반 강화학습 방법론의 간과하지 말아야 할 중요한 점은, 이 방법이 단순히 더 나은 정확도나 성능을 얻기

위한 수단을 제공하기 때문이 아니라 우리의 언어처리 과정에 대한 적절한 모델링 방법을 제시해줄 수 있기 때문이다. 그리고 본 연구를 비롯한 최근의 강화학습 관련 연구들은 특정 과제에 강화학습 알고리즘을 일원적으로 적용하는 것보다는 보다 단원적(modular)이면서도 위계적 처리 방식을 통해 전통적인 패턴인식 방법론과 강화학습 방법론을 통합적으로 활용하는 방향으로 나아가고 있다.

References

Agre, P. E., & Chapman, D. (1990). What are plans for? Robotics and autonomous systems, 6(1-2), 17–34.

Albrecht, J. E., & Clifton, C. (1998). Accessing singular antecedents in conjoined phrases. Memory & Cognition, 26(3), 599–610.

Atkeson, C. G. (1992). Memory-based approaches to approximating continuous functions. In Santa fe institute studies in the sciences of complexity-proceedings volume- (Vol. 12, pp. 521–521).

Atkeson, C. G., Moore, A., & Schaal, S. (1997). Locally weighted learning. Artificial Intelligence Review, 11(1-5), 11–73.

Baird, L. C., & Klopf, A. H. (1993). Reinforcement learning with high-dimensional, continuous actions. Wright Laboratory, Wright-Patterson Air Force Base, Tech. Rep. WL-TR-93-1147.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. IEEE transactions on systems, man, and cybernetics(5), 834–846.

Bellman, R. (1957). A Markovian decision process. Journal of Mathematics and Mechanics, 6, 679–684.

Bertsekas, D. P., & Tsitsiklis, J. N. (1989). Parallel and distributed computation: numerical methods (Vol. 23). Prentice hall Englewood Cliffs, NJ.

Bond, C., & Raehl, C. L. (2006). Adverse drug reactions in united states hospitals. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 26(5), 601–608.

Chomsky, N. (1959). A review of bf skinner's verbal behavior. Language, 35(1), 26–58.

Chomsky, N. (1969). Some empirical assumptions in modern philosophy of language. In E. Nagel, S. Morgenbesser, P. Suppes, & M. G. White (Eds.),

Philosophy, science, and method. New York: St. Martin's Press.

Clifton Jr, C., Traxler, M. J., Mohamed, M. T., Williams, R. S., Morris, R. K., & Rayner, K. (2003). The use of thematic role information in parsing: Syntactic processing autonomy revisited. Journal of Memory and Language, 49(3), 317–334.

Cook, A. E., & Myers, J. L. (2004). Processing discourse roles in scripted narratives: The influences of context and world knowledge. Journal of Memory and Language, 50(3), 268–288.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 391.

Duffy, S. A., & Keir, J. A. (2004). Violating stereotypes: Eye movements and comprehension processes when text conflicts with world knowledge. Memory & Cognition, 32(4), 551–559.

Ehrlich, K., & Rayner, K. (1983). Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. Journal of Verbal Learning and Verbal Behavior, 22(1), 75–87.

Ferreira, F., & Clifton Jr, C. (1986). The independence of syntactic processing. Journal of memory and language, 25(3), 348–368.

Frederick, S., Loewenstein, G., & O'donoghue, T. (2002). Time discounting and time preference: A critical review. Journal of economic literature, 40(2), 351–401.

Friedman, J. H., Bentley, J. L., & Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. ACM Transactions on Mathematical Software (TOMS), 3(3), 209–226.

Garrod, S., Freudenthal, D., & Boyle, E. (1994). The role of different types of anaphor in the on-line resolution of sentences in a discourse. Journal of memory and language, 33(1), 39–68.

Garrod, S., & Terras, M. (2000). The contribution of lexical and situational knowledge to resolving discourse roles: Bonding and resolution. Journal of memory and language, 42(4), 526–544.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics (pp. 174–181). Association for Computational Linguistics.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 conference on empirical methods in natural language processing, EMNLP 2014, october 25-29, 2014, doha, qatar, A meeting of sigdat, a special interest group of the ACL (pp. 1746–1751). Retrieved from `http://aclweb.org/anthology/D/D14/D14-1181.pdf`

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In Advances in neural information processing systems (pp. 3294–3302).

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188–1196).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.

Liang, C., Berant, J., Le, Q., Forbus, K. D., & Lao, N. (2017, July). Neural Symbolic Machines: Learning Semantic Parsers on Freebase with Weak Supervision. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 23–33). Vancouver, Canada: Association for Computational Linguistics. Retrieved 2017-08-03, from `http://aclweb.org/anthology/P17-1003`

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011).

Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 142–150). Association for Computational Linguistics.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. Nature, 244(5417), 522.

Marslen-Wilson, W. (1975). Sentence perception as an interactive parallel process. Science, 189(4198), 226–228.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015, February). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533. doi: 10.1038/nature14236

Mou, L., Lu, Z., Li, H., & Jin, Z. (2017). Coupling distributed and symbolic execution for natural language queries. In Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia.

Myers, J. L., Cook, A. E., Kambe, G., Mason, R. A., & O'Brien, E. J. (2000). Semantic and episodic effects on bridging inferences. Discourse Processes, 29(3), 179–199.

Myerson, J., & Green, L. (1995). Discounting of delayed rewards: Models of individual choice. Journal of the experimental analysis of behavior, 64(3), 263–276.

Neelakantan, A., Le, Q. V., Abadi, M., McCallum, A., & Amodei, D. (2017, April). Learning a Natural Language Interface with Neural Programmer. In International Conference on Learning Representations. Toulon, France. Retrieved 2017-11-30, from `https://openreview.net/forum?id=ry2YOrcge`

O'brien, E. J., Shank, D. M., Myers, J. L., & Rayner, K. (1988). Elaborative inferences during reading: Do they occur on-line? Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(3), 410.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on association for computational linguistics (p. 271).

Peng, J. (1995). Efficient memory-based dynamic programming. In Machine learning proceedings 1995 (pp. 438–446). Elsevier.

Potts, C. (2010, August). On the negativity of negation. Semantics and Linguistic Theory, 20(0), 636–659. doi: 10.3765/salt.v20i0.2565

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. Journal of verbal learning and verbal behavior, 22(3), 358–374.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613–620.

Schütze, H. (1993). Word space. In Advances in neural information processing systems (pp. 895–902).

Schütze, H. (1998). Automatic word sense discrimination. Computational linguistics, 24(1), 97–123.

Seijen, H., & Sutton, R. S. (2014, January). True Online TD(lambda). In PMLR (pp. 692–700).

Sharma, A., & Dey, S. (2012). Performance investigation of feature selection methods and sentiment lexicons for sentiment analysis. IJCA Special Issue on Advanced Computing and Communication Technologies for HPC Applications, 3, 15–20.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche,

G., ... Hassabis, D. (2016, January). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484–489. doi: 10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... Hassabis, D. (2017, October). Mastering the game of Go without human knowledge. Nature, 550(7676), 354–359. doi: 10.1038/nature24270

Singh, S. P. (1994). Learning to solve markovian decision processes (Unpublished doctoral dissertation). University of Massachusetts at Amherst.

Skinner, B. (1938). The behavior of organisms: An experimental analysis. Oxford, England: Appleton-Century.

Skinner, B. (1957). Verbal behavior. Copley Publishing Group.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., & Potts, C. (2013, 01). Recursive deep models for semantic compositionality over a sentiment treebank. , 1631, 1631-1642.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. Journal of Memory and Language, 48(3), 542–562.

Sturt, P., & Lombardo, V. (2005). Processing coordinated structures: Incrementality and connectedness. Cognitive Science, 29(2), 291–305.

Sutton, R. S. (1984). Temporal credit assignment in reinforcement learning (Ph.D. Dissertation). University of Massachusetts, Amherst, MA.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. Machine learning, 3(1), 9–44.

Sutton, R. S. (1996). Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), Advances in Neural Information Processing Systems 8 (pp. 1038–1044). MIT Press.

Sutton, R. S., & Barto, A. G. (1981). Toward a mordern theory of adaptive networks: Expectation and prediction. Psychological Review, 88, 135–

170.

Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction. MIT press Cambridge.

Szepesvári, C. (2010). Algorithms for reinforcement learning. Synthesis lectures on artificial intelligence and machine learning, 4(1), 1–103.

Taboada, M., Anthony, C., & Voll, K. (2006). Creating semantic orientation dictionaries. In Proceedings of 5th international conference on language resources and evaluation (p. 427-432). LREC.

Tadepalli, P., & Ok, D. (1996). Scaling up average reward reinforcement learning by approximating the domain models and the value function. In Icml (pp. 471–479).

Tesauro, G. (1995). Td-gammon: A self-teaching backgammon program. In Applications of neural networks (pp. 267–285). Springer.

Thorndike, E. (1911). Animal intelligence: Experimental studies. New York, NY: Macmillan.

Tolman, E. (1948). Cognitive maps in rats and men. Psychological Review, 55(4), 189–208.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417–424). Association for Computational Linguistics.

Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems, 24(7), 1024–1032.

Van Gompel, R. P., & Majid, A. (2004). Antecedent frequency effects during the processing of pronouns. Cognition, 90(3), 255–264.

Van Seijen, H., Mahmood, A. R., Pilarski, P. M., Machado, M. C., & Sutton, R. S. (2016). True online temporal-difference learning. Journal of

Machine Learning Research, 17(145), 1–40.

Watkins, C. J., & Dayan, P. (1992). Q-learning. Machine learning, 8(3-4), 279–292.

Wiebe, J. (2000). Learning subjective adjectives from corpora. In Proceedings of 17th national conference on artificial intelligence (p. 0). AAAI.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 8(3-4), 229–256.

Witten, I. H. (1977). An adaptive optimal controller for discrete-time markov environments. Information and control, 34(4), 286–295.

Yang, Y., & Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. In Proceedings of the Fourteenth International Conference on Machine Learning (pp. 412–420). San Francisco, CA, USA.

Zagibalov, T., & Carroll, J. (2008). Automatic seed word selection for unsupervised sentiment classification of chinese text. In Proceedings of the 22nd international conference on computational linguistics-volume 1 (pp. 1073–1080).

Zhong, V., Xiong, C., & Socher, R. (2017, August). Seq2sql: Generating Structured Queries from Natural Language using Reinforcement Learning. arXiv:1709.00103 [cs]. Retrieved 2017-11-15, from `http://arxiv.org/abs/1709.00103` (arXiv: 1709.00103)

Abstract

A Study of Measuring Semantic Orientation of Words and State Values of Statements by Reinforcement Learning

Youngsam Kim

Cognitive Science Program

The Graduate School

Seoul National University

In this paper, reinforcement learning indicates a kind of machine learning methods, which its learning is achieved by maximizing the rewards from an environment. The fundamental difference of this approach from various 'supervised learning methods' is that it prompts the learning agent to try different actions for a problem and find the best solution itself, without giving an explicit direction. In this process, actions of the agent are taken to maximize not only immediate rewards, but also successive profits.

The main objective of this paper is to process written texts in an incremental, continuous fashion with the methods of reinforcement learning. However, there are too many fields or tasks in computational linguistics/natural language processing, and thus only two extensible tasks are investigated in this dissertation. The first task is to estimate semantic orientation of words in texts and the second one is to evaluate a nursing statement on patients in respect to

Adverse Drug Reaction. Both tasks are attempted to be solved using temporal-difference learning, a well-established algorithm in reinforcement learning.

Much psychological research found that our language processing mechanism is inherently automatic and incremental. Thus, applying reinforcement learning to realistic problems in natural language processing seems to lead into a cognitive approach. And this thesis explores the possibility of utilizing the psychological algorithm for the questions of natural language processing.

Keywords: reinforcement learning, natural language processing, semantic orientation, state values, incremental processing
Student Number: 2009-30810