신경망을 이용한 반자동 구문분석 말뭉치 구축도구

임준호[0] 곽용재 박소영 임해창
고려대학교 컴퓨터학과
{jhlim[0], yjkwak, ssoya, rim}@nlp.korea.ac.kr

Semi-Automatic Tree Annotating Workbench Using Neural-Networks

Joon-Ho Lim[0], Yong-Jae Kwak, So-Young Park, Hae-Chang Rim
Dept. of Computer Science, Korea University

요 약

구문분석 말뭉치는 통계적 구문분석 분야의 필수적인 항목으로 많은 유용성을 가지지만, 말뭉치를 구축할 때, 막대한 시간과 비용이 요구되기 때문에 구축자의 수작업을 감소시키는 방법에 대한 연구가 필요하다. 본 논문에서는 대량의 신뢰도 있는 구문분석 말뭉치를 구축하기 위해 신경망을 사용하는 반자동 구문분석 말뭉치 구축도구에 대해서 설명한다. 개발된 도구는 구문패턴 추출, 신경망 학습, 반자동 구축의 세 단계로 구성된다. 구문패턴 추출 단계에서는 사용자가 정의한 자질집합을 사용하여 기존에 구축된 말뭉치에서 구문패턴들을 추출하고, 신경망 학습의 단계에서는 추출된 구문패턴들을 사용하여 신경망을 학습한다. 그리고, 반자동 구축 단계에서는 학습된 신경망을 사용하여 반자동으로 구문분석 말뭉치를 구축한다. 본 논문에서 제안하는 방법은 다양한 자질집합을 조합하여 사용할 수 있고, 학습을 사용하기 때문에 학습집합에 나타나지 않은 경우에 대해서도 합리적인 결정을 내릴 수 있다. 소량의 구문분석 말뭉치를 대상으로 실험한 결과, 본 논문에서 제안하는 방법이 약 42.5%의 수작업 횟수 감소율을 보였음을 알 수 있었다.

1. 서 론

말뭉치는 실세계에서 사람들이 사용하는 언어를 기계 가독 형태(machine-readable form)로 저장하여 놓은 언어 데이터를 말하고, 구문분석 말뭉치는 각 문장에 대하여 구문구조를 부착한 말뭉치를 말한다[1]. 이와 같은 구문분석 말뭉치를 사용하면 특정 규칙의 발생 확률, 두 단어 사이의 의존 확률, 등 구문분석을 수행하는데 필요한 여러 가지 정보를 자동으로 추출할 수 있다. 그렇기 때문에, 구문분석을 수행하는 자연어처리 분야에서는 구문분석 말뭉치가 필수적으로 사용되고 있다.

구문분석 말뭉치는 이와 같이 많은 유용한 정보를 제공하지만, 이를 구축하기 위해서는 구축자의 막대한 노력과 시간이 요구된다. 그렇기 때문에, 구문분석 말뭉치를 구축할 때 구축자의 수작업을 감소시켜 줄 수 있는 방법에 대한 연구가 필요하다.

구문분석 말뭉치를 구축할 때 구축자의 수작업을 감소시키기 위하여 본 논문에서는 신경망을 이용한 구문패턴의 학습, 적용 방법을 제안한다. 2장에서는 효과적인 말뭉치 구축을 위한 기존연구들을 살펴보고, 3장에서는 본 논문에서 제안하는 구문패턴 추출, 신경망 학습, 반자동 구축 방법에 대해서 설명한다. 마지막으로, 4장에서는 실험을 통하여 본 논문에서 제안하는 방법이 기존의 방법들보다 얼마나 더 수작업을 감소시킬 수 있는지를 알아본다.

2. 기존 연구

구문분석 말뭉치를 구축할 때 구축자의 수작업을 감소시키기 위한 연구는 크게 두 가지로 나눠 볼 수 있다. 첫 번째는 말뭉치 구축을 도와주기 위한 규칙들을 구축도구 안에 내장시키는 방법([2,3])이고, 두 번째는 기존에 구축된 말뭉치를 이용하여 말뭉치 구축을 도와주는 규칙을 추출하는 방법([4])이다. 첫 번째 방법은 PennTreeBank([2]), STEP2000([5]) 등을 구축할 때 사용된 방법으로, 한 문장이 입력되었을 때, 입력 문장 중 중의성이 없는 부분에 대해서 규칙을 사용하여 부분적인 구문구조를 할당하여 준다. 이 방법은 문법전문가가 수동으로 규칙을 작성하였기 때문에, 규칙에 대한 신뢰도가 높다는 장점을 가진다. 하지만, 규칙이 한 번 정해진 이후에는 수정이 힘들고, 규칙을 만드는데 사용된 자질집합을 변경하기가 힘들다는 단점을 가지고 있다. 그리고, 부분 구문구조가 부착된 이후의 모든 작업을 구축자가 부담해야 하기 때문에, 비효율적인 부분이 있다.

두 번째 방법은 [4]에서 사용된 방법으로, 기존에 구축된 말뭉치에서 필요한 정보를 추출하여 사용하는 방법이다. 두 번째 방법은 첫 번째 방법과 비교하여, 규칙을 자동으로 추출하기 때문에 첫 번째 방법보다 신뢰도가 떨어지는 규칙을 사용할 수도 있다는 단점이 있다. 하지만, 기존에 구축된 말뭉치에서 규칙을 추출하기 때문에 다양한 자질집합을 사용하기가 용이하고, 구문구조를 부착하는 중 언제라도 규칙의 도움을 받을 수 있도록 작성되었다는 장점을 가지고 있다. 그리고, 기존에 구축된 말뭉치의 양이 증가할수록 더 쉽게 구문구조를 부착할 수 있다는 장점을 가지고 있다.

하지만, 기존에 연구되었던 [4]의 규칙 추출 방법은 몇 가지 문제점을 가지고 있다. 기존 방법은 실험집합에 나타났던 각각의 개별적인 구문패턴들 자체에 대해서만 통계적 신뢰도를 사용하여 선별하였기 때문에 학습집합에서 나타나지 않았던 구문패턴에 대해서는 아무런 판단을 할 수가 없었고, 입력 자체의 오류(noise)에 대해서도 견고하게(robust) 동작하지 못하였다. 본 논문에서는 신경망을 사용하여 구문패턴을 학습함으로써 이와 같은 단점들을 해결하고자 한다.

3. 신경망을 이용한 반자동 구문분석 말뭉치 구축도구

본 논문에서 구현한 반자동 말뭉치 구축도구는 이진 구구조 문법을 기초로 사용하고, 구문구조를 부착하는데 묶기/이동(Reduce/Shift)의 LR 연산을 사용한다.

전체적인 작업 과정은 그림1과 같다. 우선, 기존에 구축되어 있는 말뭉치가 없는 경우, 수동으로 구문분석 말뭉치를 구축한다. 그리고, 수동으로 구축된 말뭉치가 쌓이면 사용자가 정의한 자질집합에 따라 구문패턴들을 추출하고, 추출된 구문패턴을 사용하여 신경망을 학습한다. 이렇게 학습된 신경망이 다음 말뭉치를 구축할 때, 반자동으로 구문분석을 수행할 수 있도록 도와준다.

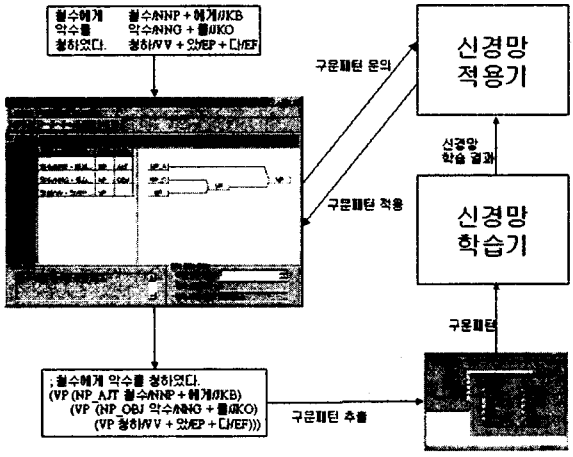

그림 1 시스템 구성도

3.1 구문패턴 추출

기존에 구축된 말뭉치에서 구문패턴을 추출하는 방법은 다음과 같다. 예를 들어, 그림2와 같은 구문분석 결과가 있을 때, 좌/우 구문태그와 기능 태그, 품사열을 자질로 사용하여 구문패턴을 추출한다면 그림 3과 같은 구문패턴들이 추출된다.

```
(VP (NP_MOD  철수/NNP + 에게/JKB)
    (VP  (NP_OBJ  악수/NNG + 를/JKO)
         (VP  청하 /VV + 았 /EP + 다 /EF)))
```

그림 2 구문패턴 추출 입력 예제

결과	좌 구문-기능 태그	좌 중심어 품사열	우 구문-기능 태그	우 중심어 품사열	Notes
이동	NP_MOD	NNP +JKB	NP_OBJ	NNG +JKO	철수에게 악수를
묶기	NP_OBJ	NNG +JKO	VP	VV +EP +EF	악수를 청하였다

그림 3 구문패턴 추출 출력 예제

본 도구에서 추출할 수 있는 자질들은 다음과 같이 정했다.

- 좌/우 구문기능 범주
- 좌/우 중심어절의 품사열
- 좌/우 중심어절의 어휘열
- 좌/우 어절수
- 좌/우 격정보
- 좌/우 외부문맥

3.2 신경망 학습

본 논문에서는 Sigmoid 유닛으로 신경망을 구성하고, 3.1절에서 추출한 구문패턴들을 역전파 알고리즘을 (Back-Propagation Algorithm) 사용하여 학습하였다[6].

1개의 입력 유닛은 1개의 자질에 대응되기 때문에, 입력계층(Input Layer)의 입력 유닛의 개수는 자질들의 개수와 같다. Hidden Layer는 기본적으로 3개의 유닛을 사용하여 입력 계층에서 출력된 값을 표현한다. 출력 계층은 2개의 유닛을 가지고, 각각의 유닛이 묶기와 이동에 대한 신뢰도를 나타낸다.

구문패턴의 값은 구문태그나 품사열과 같은 기호(symbol) 값이기 때문에, 이를 사용하여 신경망을 학습하기 위해서는 입력 값을 다른 형태로 변환하여 사용해야 한다. 본 논문에서는 신경망의 입력 값으로 불린 값의 벡터(boolean-valued vector)를 사용한다. 예를 들어, 총 가능한 태그의 개수가 41개라면, 입력 유닛은 41차원의 벡터 값을 입력으로 받는다. 그리고, 41개 중 NP_MOD가 19번째에 해당하는 태그 값이라면, NP_MOD라는 태그는 19번째 값만 1이고 나머지는 모두 0인 태그 값으로 신경망에 입력된다. 만약, 입력이 중심어절의 품사열/어휘열과 같이 여러 개의 태그 값을 가지는 경우라면, 각 태그에 해당하는 부분만 1의 값을 가지고 나머지는 0의 값을 가지는 벡터를 입력으로 가지게 한다.

3.3 반자동 구축

본 도구는 구문구조를 반자동으로 부착하기 위해서 3.2절에서 학습한 신경망을 사용한다. 반자동 구축도구는 신경망 학습 결과와 사용자가 정의한 임계값을 입력으로 받고, 다음과 같은 과정을 거쳐서 반자동 구문분석을 수행한다.

(가) 문장을 입력 받는다.
(나) 현재 상태를 구문패턴으로 만들고, 그 값을 신경망에 대입하고, 결과 unit의 값들을 받아온다.
(다) 두 결과 unit사이의 차이 값이 사용자가 정의한 임계값보다 높다면, 높은 값에 해당하는 연산을 적용하고 (나)의 단계로 돌아간다.
(라) 차이 값이 사용자 정의 임계값보다 작다면, 구축자가 지금까지 신경망에서 적용한 결과가 맞는지 확인한다.
(마) 적용한 결과가 틀리다면 연산을 취소하고 올바른 연산을 수행시켜 준다. 그리고, (나)의 과정으로 돌아간다.
(바) 적용된 결과가 맞다면 구축자가 올바른 묶기/이동 연산을 수행시켜주고, (나)의 과정으로 돌아간다.
(사) 문장에 대해서 올바른 구문구조가 부착되었다면 끝내도록 한다.

4. 실험 및 평가

실험은 신경망을 사용하여 구문패턴을 학습하는 방법이 구축자의 수작업 감소에 얼마나 도움이 될 수 있는가를 알아보는 목적으로 수행되었다.

실험에 사용된 말뭉치는 [7]에서 수동으로 구축한 말뭉치로서, [7]에서 정의한 태그집합과 구문구조를 따라서 분석한 말뭉치이다. 말뭉치는 총 1,397문장으로, 이 중의 90%(1,256문장)를 학습집합으로 사용하고, 10%(141문장)를 실험집합으로 사용하였다. 그리고, 말뭉치의 양이 작기때문에, 표 1에 명시된 세 가지 자질집합에 대해서만 실험을 수행하였고, 신경망의 결과값을 적용하는 기준이 되는 임계값은 0.5를 사용하였다.

표 1 자질집합

자질집합	자질 조합 (좌/우)
A	구문-기능 태그
B	구문-기능 태그+중심어절의 품사열
C	구문-기능 태그+중심어절의 품사열+어절수

신경망을 이용한 반자동 구문분석이 얼마나 구축자의 수작업을 감소시키는지를 알기 위해서는 적용된 결과가 맞는지 틀린지 확인하는 작업까지 고려해야한다. 하지만, 이는 수량적으로 측정하기 힘들기 때문에, 본 연구에서는 이를 고려하지 않았고, 패턴 정확률, 패턴 재현율, 수작업 횟수 감소율을 사용하여 성능을 평가하였다. 패턴 정확률은 적용된 구문패턴이 얼마나 정확했는가에 대한 성능을 나타내고, 패턴 재현율은 전체 구문구조에 대해 정답 패턴이 적용된 비율에 대한 성능을 나타낸다. 수작업 횟수 감소율은 적용결과를 틀렸을 경우 이를 취소하는 작업까지 고려하여 실제 구축자의 수작업 횟수가 감소된 비율을 계산한 값이다. 각 평가 방법에 대한 수식은 다음과 같다.

$$\text{패턴정확률} = \frac{\text{맞은적용수}}{\text{맞은적용수} + \text{틀린적용수}}$$

$$\text{패턴재현율} = \frac{\text{맞은적용수}}{\text{수동 구축도구에서의 수작업수}}$$

$$\text{수작업 횟수 감소율} = \frac{\text{맞은적용수} - \text{틀린적용수}}{\text{수동 구축도구에서의 수작업수}}$$

표 2는 신경망을 사용하여 구문패턴을 학습한 실험 결과와 신뢰도기반의 구문패턴 선별([4]) 방법을 실험한 결과이다. 실험 결과를 분석하여 보면 다음과 같다. 자질집합 A의 경우 신뢰도 기반 방법은 맞은 적용수에 비해서 틀린 적용수가 더 많았기 때문에 전체적인 수작업 횟수 감소율이 음수의 값으로 나타났고, 신경망을 사용하여 학습한 경우는 신뢰도 기반 방법에 비해서 틀린 적용수가 크게 감소하여 전체적인 수작업 횟수 감소율이 70% 가량 증가하게 되었다. 자질집합 B의 경우 역시 맞은 적용수가 증가하고, 틀린 적용수가 감소하였는데, 이는 학습집합에 나타나지 않았던 구문패턴에 대해서도 올바른 판단을 하였기 때문이고, 그 결과 전체적인 수작업 횟수 감소율이 42.5%를 나타냄으로써 가장 높은 수작업 감소 효과를 나타낼 수 있는 것으로 나타났다. 자질집합 C의 경우, 맞은 적용수가 증가하여 패턴 재현율이 증가하였지만, 틀린 적용수도 증가하여 패턴 정확률의 값은 조금 떨어지게 되었고, 수작업 횟수 감소율은 10% 정도 증가하였다.

표 2 실험 결과

구문패턴	자질 집합	패턴 정확률	패턴 재현율	수작업 횟수 감소율
신경망 학습	A	**65.7%**	**54.8%**	**26.2%**
	B	**75.8%**	**62.4%**	**42.5%**
	C	**76.1%**	**60.2%**	**41.3%**
신뢰도 기반 선별 ([4])	A	33.9%	60.0%	-56.5%
	B	69.8%	52.0%	29.6%
	C	81.7%	39.5%	30.6%

5. 결론 및 향후 연구

본 논문에서는 대량의 신뢰도 있는 구문분석 말뭉치를 구축하기 위해 신경망을 사용하는 반자동 구문분석 말뭉치 구축도구에 대해서 설명하였다. 개발된 도구는 구문패턴 추출, 신경망 학습, 반자동 구축의 세 단계로 구성되고, 신경망 학습을 사용하였기 때문에, 학습집합에 나타나지 않았던 구문패턴에 대해서도 올바른 판단을 할 수 있었다. 그 결과 기존의 신뢰도 기반 구문패턴 추출 방법보다 더 좋은 성능을 나타냈다.

향후 작업으로는 더 정확한 구문분석 말뭉치를 구축할 수 있도록 자질잡합과 패턴 정확률, 패턴 재현율 사이의 관계를 살펴보고, 다른 기계학습 기법들을 적용하여 볼 것이다.

참고 문헌

[1] 류원호, 이상주, 임해창, "어휘 문맥 의존 규칙과 통계 모델을 이용한 한국어 품사 부착 말뭉치 구축 도구", 고려대학교 석사학위 논문, 1999.

[2] Mitchell P. Marcus, B. Santorini, and M. A. Marcinkiewicz, " Building a large annotated corpus of English : the Penn Treebank", Computational Linguistics, Vol.19, No.2, pp.313-330, 1993.

[3] 장병규, 이공주, 김길창, "대량의 한국어 구문 트리 태깅 코퍼스 구축을 위한 구문 트리 태깅 워크 벤치의 설계 및 구현", 제 9회 한글 및 한국어 정보처리 학술 발표 논문집, pp.421-429, 1997.

[4] 임준호, 박소영, 곽용재, 임해창, 김의수, 강범모, "구문패턴을 이용한 반자동 구문분석 말뭉치 구축도구", 제 14회 한글및 한국어 정보처리학회, pp.343~350, 2002.

[5] 이공주, 김재훈, 장병규, 최기선, 김길창, "한국어 구문 트리태깅 코퍼스 작성을 위한 한국어 구문태그". CS/TR-96-102, KAIST, 1996.

[6] Tom M. Mitchell. "Machine Learning", McGraw-Hill, 1997.

[7] 김흥규 외, "1장4절 구문분석 말뭉치 개발", 21세기 세종계획 국어기초자료 구축 보고서, pp.199~252, 2002