구문패턴을 이용한 반자동 구문분석 말뭉치 구축도구

임준호○ 박소영+ 곽용재+ 임해창+ 김의수++ 강범모++

○, + : 고려대학교 컴퓨터학과 자연어처리 연구실

{ jhlim, ssoya, yjkwak, rim }@nlp.korea.ac.kr

++ : 고려대학교 민족문화연구원

usk2000@orgio.net, bmkang@korea.ac.kr

요약

본 논문에서는 구문패턴을 이용한 반자동 구문분석 말뭉치 구축도구를 제안한다. 일반적으로 구문분석 말뭉치를 구축하는 작업은 문법전문가의 많은 시간과 노력을 필요로 하고 있다. 본 논문은 구문분석 말뭉치를 구축할 때 수작업을 감소시켜 줄 수 있는 도구를 개발하기 위하여, 사용자가 정의하는 자질집합과 신뢰도를 바탕으로 구문패턴을 자동 추출하고 적용하는 방법을 제안한다. 소량의 말뭉치에서 실험한 결과, 구문패턴의 사용은 30%정도의 수작업을 감소시킬 수 있는 것으로 나타났다.

1. 서론

말뭉치란 실세계에서 사람들이 사용하는 자연어를 기계가 읽을 수 있는 형태로 컴퓨터에 저장해 놓은 언어 정보를 말한다[1]. 이와 같은 말뭉치에는 다른 부가정보가 없이 단어나 문장만이 있는 원시 말뭉치, 어절을 분석하여 각 형태소에 품사를 부착한 형태소 분석 말뭉치, 문장을 분석하여 문장의 구조를 부착한 구문분석 말뭉치, 각 단어의 의미를 분석한 의미분석 말뭉치 등이 있다.

언어를 분석하는 작업은 언어학적인 이론이나 가설을 바탕으로 수행할 수 있지만, 이런 이론이나 가설은 언어학자의 주관적인 문법적 판단에 영향을 받을 수 있으며, 시대의 흐름에 따른 언어의 변화를 적절히 반영할 수 없다. 이들 말뭉치는 이런 주관적일 수 있는 언어가설을 검증할 수 있는 객관적인 정보를 제공한다[2]. 이런 이유로, 말뭉치는 언어학뿐만 아니라 자연어 처리의 여러 분야에서 유용하게 사용되고 있으며 이는 구문분석 분야에서도 마찬가지이다. 즉, 구문분석 말뭉치에서 어떤 단어나 품사의 발생 빈도, 문법의 사용 빈도, 단어 혹은 품사들의 상호정보 등을 추출하고 이를 활용하면 효과적으로 구문분석을 수행할 수 있다[3,4]. 이를 위해서는 구문분석 말뭉치의 구축이 선행되어야 한다.

구문분석 말뭉치를 구축하는데 있어서 중요한 점은 그 크기가 충분히 커야하고, 포함되어 있는 정보가 정확해야 한다는 것이다. 그러나, 정확한 말뭉치의 구축은 아직 컴퓨터가 자동으로 처리할 수 없는 문제이므로, 사람이 각 문장을 보고 수동으로 구문분석을 수행한다. 이처럼 사람이 수동으로 말뭉치를

※ 본 논문은 2002년도 21세기 세종계획의 지원을 받아 진행된 연구의 일환으로 쓰여졌음

구축하는 작업은 많은 시간과 인력을 필요로 하는 작업이고, 그렇기 때문에 사람의 수작업을 줄여줄 수 있는 방법이 필요하다.

본 논문에서는 구문분석 말뭉치를 구축함에 있어서 사람의 수작업을 줄여줄 수 있는 구문패턴을 이용하는 방법을 제안하고자 한다. 2장에서는 기존의 말뭉치 구축도구에서 수작업 감소를 위해 사용한 방법들의 장단점을 비교 분석한다. 그리고, 3장에서는 제안하는 말뭉치 구축 도구에 대해 설명하고, 4장에서는 제안하는 방법의 수작업 감소효과를 실험을 통해 평가한다. 마지막으로 5장에서는 본 논문에서 제안한 방법에 대해서 결론을 내리고, 향후연구를 제시한다.

2. 기존 연구

국내에서 자연어 처리를 위해 사용되는 대표적인 구문분석 말뭉치로는 PennTreeBank [5], STEP2000 구문분석 말뭉치[6] 등이 있으며, 이들은 수작업 감소를 위해 몇 가지 방법을 이용한다[5,7,8]. 이 장에서는 말뭉치를 구축하는 도구의 수작업 감소 비율, 규칙의 자동추출 여부, 사용자의 개입위치, 규칙의 이용정보를 고려하여 각 방법들을 비교한다.

PennTreeBank는 WSJ 신문기사와 ATIS의 약 4만 8천 문장을 분석한 말뭉치인데, Fidditch[9]라는 구문분석 도구를 사용하여 구축하였다[5]. Fidditch는 입력되는 문장에 대해서 문장 성분 정보만을 고려하여 중의성 없는 부분만 구문구조를 부착하고, 중의성이 있는 부분에 대해서는 구문구조를 부착하지 않고 사용자의 결정에 따른다. 이 방법은 규칙을 이용하므로 사용할 수 있는 정보가 고정되어 있게 되고, 문장이 입력될 때 한번만 부분 구문구조를 부착하기 때문에 구문구조가 부착된 이후에는 모든 일을 사람이 부담해야 한다는 단점이 있다. 하지만, 본 연구에서 제안하는 방법은 규칙의 자질집합을 선택할 수 있고, 구문분석을 수행하는 중간에 언제라도 규칙의 도움을 받을 수 있다.

STEP2000 구문분석 말뭉치[6]는 1만 문장을 분석한 말뭉치인데, 구문구조 부착과 검증의 두 단계로 동작하는 도구를 사용하여 구축하였다[7]. 첫째, 부분 구문구조 부착 단계는 전문가가 손으로 뽑은 규칙을 가지고 입력 문장에 대해서 최장길이 검색을 하여 해당하는 부분을 자동적으로 부분 구문구조로 대체하여 주는 것이다. 둘째, 검증 단계는 먼저 구축된 말뭉치에서 일정 빈도 이상의 것을 추출하여 사용자가 맞다고 검증한 규칙들과 현재 사용자가 구문구조를 부착하는 규칙을 비교하는 작업이다. 이 도구는 이와 같이 부착과 검증 과정을 거쳐서 수작업을 감소시켰고, 정확한 말뭉치를 구축할 수 있도록 하였다. 하지만, 이 방법은 규칙을 수동으로 추출하고, 부분 구문구조가 만들어진 이후에는 규칙을 사용할 수 없고, 규칙을 구성하는 자질집합을 도구 내부에 고정시켜 두었기 때문에 차후 변경이 불가능하다는 문제점이 있다. 이에 반해서, 본 연구에서 제안하는 방법은 말뭉치에서 원하는 자질을 사용하는 규칙들을 자동으로 추출하고, 이를 통계적 신뢰도를 사용하여 검증한다. 그리고, 구문구조가 부착되는 중간에 계속해서 규칙의 도움을 받을 수 있도록 개발되었다.

[4]는 93년 동아일보 사설에서 추출한 550문장에 대해 구문분석을 수행하는 데에 묶기/이동(Reduce/Shift)으로 구성되는 LR연산을 사용하였다. 이 작업은 제한된 길이의 문맥 품사열을 획득하는 과정과 적용하는 과정을 분리하여 사용한다. 문맥 품사열을 획득하는 과정에서는 묶임/이동을 수행하게 되는 노드의 문맥 품사열을 기록한다. 그리고, 적용하는 과정에서는 묶임/이동을 판단해야 되는 후보 노드의 문맥 품사열을 살펴보고 기존에 저장된 것이 있다면 이를 적용하여 준다. 이 방법의 단점은 규칙을 뽑는데 문맥 품사열만을 사용하였기 때문에, 여러 가지 자질을 사용하지 못하고 자질이 고정되어 있다는 점이다. 하지만, 본 연구에서 제안하는 방법은 사

용자가 원하는 대로 여러 개의 자질 가운데 필요한 자질을 선택해서 구문패턴들을 구성할 수 있다.

3. 반자동 구문분석 말뭉치 구축도구

이 장에서는 반자동 구문분석 말뭉치 구축도구가 어떻게 작동하는 지를 설명한다. 즉, 구문패턴을 추출하고 말뭉치 구축도구에 활용하는 방법을 자세히 기술한다.

3.1 개요

본 논문에서 이미 개발하여 사용하고 있는 수동 구문분석 말뭉치 구축도구[10]는 입력문장에 구문분석을 수행하는 데에 묶기/이동의 LR연산을 사용한다. 이 도구의 입력은 형태소 분석결과이고, 출력은 구문부착 결과이다. 구문분석 형식은 이진 구구조 형식을 따른다[10].

그림 1은 반자동 구문분석 말뭉치 구축도구의 전체적인 과정을 나타낸다. 수동 구문분석 말뭉치 구축도구를 반자동 구문분석 말뭉치 구축도구로 개선하기 위해서 본 논문에서는 구문패턴을 사용한다. 사용자가 자질집합과 신뢰도를 선택하면, 추출된 구문패턴 후보에 대해 신뢰도를 검사하고, 주어진 신뢰도를 만족하는 후보를 구문패턴으로 선별한다.

3.2 구문패턴 후보 추출

제안하는 방법은 기존의 말뭉치에서 필요한 정보를 추출하여 말뭉치를 구축하는데 활용한다. 이렇게 필요한 정보를 추출하는 단계가 구문패턴의 후보를 추출하는 단계이다.

묶기/이동 연산을 사용할 때, 주어진 구문분석 말뭉치에서 구문패턴을 추출하는 과정은 결정적이다. 즉, 그림2와 같은 문장이 있을 때, 묶기/이동의 연산을 사용하여 구문분석을 수행한다면, 이동-묶기-묶기의 연산을 수행할 때 그림 2와 같은 구문분석 결과가 생성된다. 하지만, 연산 순서를 다르게 하여 적용한다면 그림 2와 같은 구문분석 결과를

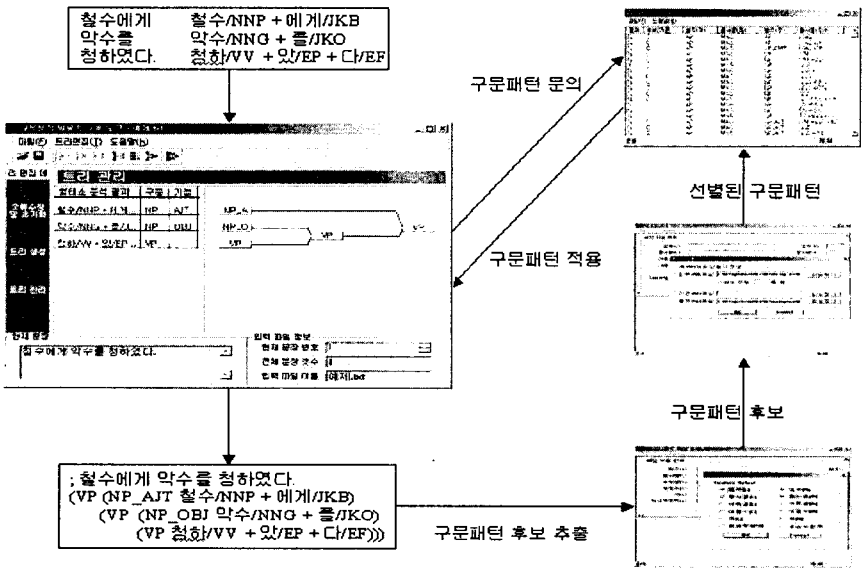

그림 1 반자동 구문분석 말뭉치 구축도구

생성할 수 없게 된다. 구문패턴 후보를 추출하는 단계가 이와 같이 결정적이기 때문에, 수동으로 구축한 모든 말뭉치에 대해서 일괄적으로 구문패턴 후보를 추출할 수 있다.

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

그림 2 입력문장 예제

구문패턴 후보를 추출하기 위해서는 먼저 어떤 자질을 구문패턴에 포함시킬지를 결정해야 한다. 즉, 어떤 자질을 고려하여 구문패턴을 만들었을 때, 유용한 구문패턴을 만들 수 있는지를 결정해야 한다. 이 시스템에서 선택할 수 있는 자질들은 다음과 같다.

- 좌우 구문-기능 범주
- 좌우 중심어절의 품사열
- 좌우 중심어절의 어휘열
- 좌우 어절수
- 좌우 격정보
- 좌우 외부문맥

도구에서 자질집합을 선택하는 장면은 그림3과 같다. 예를 들어, 그림 2의 문장에서 좌우 구문-기능 범주와 중심어의 품사열을 자질로 고려하여 구문패턴 후보를 추출한다면, "철수에게 악수를"과 "악수를 청하였다"에 해당하는 구문패턴 후보는 그림 4에 나타난 두 가지 구문패턴 후보가 된다.

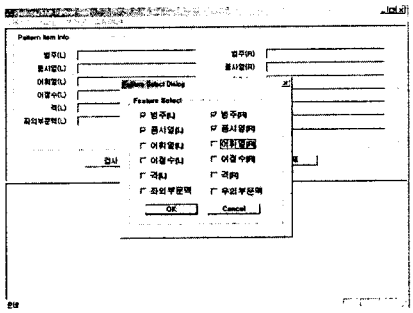

그림 3 자질집합 선택

3.3 통계기반 구문패턴 자동추출

위의 3.2절에서 추출한 구문패턴 후보들은 말뭉치에서 한 번 이상 나타난 구문패턴들을 모두 추출한 것이다. 이렇게 추출된 구문패턴 후보들은, 같은 자질 값을 갖는 경우에 대해서도 서로 다른 결과를 내기도 하고, 전체 말뭉치 중 한번만 나타나기도 한다. 결과적으로, 3.2절에서 추출한 모든 구문패턴 후보를 그대로 다음 말뭉치를 구축하는데 사용하는 것은 무리가 있다. 그래서 본 연구에서는 통계적 가설 검정을 사용하여, 일정 수준 이상의 신뢰도를 가지는 구문패턴 후보에 대해서만 구문패턴으로 사용하도록 한다.

통계적 가설검정을 하는 데에는 누적이항

결과	좌구문-기능	좌품사열	우구문-기능	우품사열	비고
이동	NP_MOD	NNP +JKB	NP_OBJ	NNG +JKO	철수에게 악수를
이동	NP_MOD	NNP +JKB	NP_OBJ	NNG +JKO	
이동	NP_MOD	NNP +JKB	NP_OBJ	NNG +JKO	
묶기	NP_OBJ	NNG +JKO	VP	VV +EP +EF	악수를 청하였다
묶기	NP_OBJ	NNG +JKO	VP	VV +EP +EF	
묶기	NP_OBJ	NNG +JKO	VP	VV +EP +EF	
묶기	NP_OBJ	NNG +JKO	VP	VV +EP +EF	
이동	NP_OBJ	NNG +JKO	VP	VV +EP +EF	

그림 4 추출된 구문패턴 후보

분포를 사용한다. 그리고, 임의의 구문패턴 후보가 구문패턴이 아닐 확률을 우연율로 정의하고, 누적이항분포를 사용한 신뢰도를 다음과 같이 정의한다[11].

$$\begin{aligned}\text{신뢰도} &= 1 - \text{우연율} \\ &= 1 - P(X \geq r) \\ &= 1 - \sum_{i=r}^{n} \frac{n!}{i!(n-i)!} \times p^i \times (1-p)^{n-i}\end{aligned}$$

위의 식에서, n은 구문패턴이 나타난 횟수이고, r은 묶기/이동이 선택된 횟수이다. p는 임의의 패턴이 묶기나 이동을 선택될 확률이므로 0.5로 가정한다. 즉, 우연율과 신뢰도는 반비례 관계를 가지게 된다.

예를 들어, 그림 2의 문장에서 "철수에게 악수를"과 "악수를 청하였다."에 대한 구문패턴 후보를 추출한 것이 그림 4와 같이 추출되었다면, 이 구문패턴 후보에 대한 가설검정은 다음과 같이 이뤄진다.

1) 철수에게 악수를

$$\begin{aligned}\text{신뢰도} &= 1 - \text{우연율} \\ &= 1 - P(X \geq 3) \\ &= 1 - \sum_{i=3}^{3} \frac{3!}{i!(3-i)!} \times \frac{1}{2}^{i} \times \frac{1}{2}^{3-i} \\ &= 1 - 0.125 \\ &= 0.875\end{aligned}$$

2) 악수를 청하였다.

$$\begin{aligned}\text{신뢰도} &= 1 - \text{우연율} \\ &= 1 - P(X \geq 4) \\ &= 1 - \sum_{i=4}^{5} \frac{5!}{i!(5-i)!} \times \frac{1}{2}^{i} \times \frac{1}{2}^{5-i} \\ &= 1 - 0.187 \\ &= 0.813\end{aligned}$$

신뢰도 85%를 기준으로 구문패턴을 추출한다면, "철수에게 악수를"에 해당하는 구문패턴 후보는 구문패턴으로 추출되고, "악수를 청하였다."에 해당하는 구문패턴 후보는 구문패턴으로 추출되지 않는다.

3.3 구문패턴 적용과정

위의 과정을 거친 구문패턴은 이제 확률적으로도 높은 신뢰도를 가지는 것들이다. 이 구문패턴을 사용하여 실제 구문분석을 수행하는 작업은 그림 5와 같은 과정을 거쳐 이뤄진다.

a) 문장을 입력받는다.
b) 현재 상태에서 적용할 수 있는 구문패턴이 있는지 검사한다.
c) 구문패턴이 있다면, 그 구문패턴을 적용하고 b)의 과정으로 돌아간다.
d) 구문패턴이 없다면, 지금까지 적용된 구문패턴들이 맞는지 사람이 확인한다.
e) 적용된 구문패턴이 틀리다면, 연산을 취소하고 올바른 연산을 수행시켜 준다. 그리고, b)의 과정으로 돌아간다.
f) 적용된 구문패턴이 맞다면, 사람이 올바른 묶기/이동 연산을 수행시켜주고, b)의 과정으로 돌아간다.
g) 문장에 대해서 올바른 구문구조가 부착되었다면 끝내도록 한다.

그림 5 패턴 적용과정

4. 실험 및 평가

실험은 패턴의 정확율, 재현율, 수작업 감소율을 알아보는 목적으로 수행되었다.

4.1 실험 환경

사용한 말뭉치의 분석단위는 어절단위이다. 실험에 사용한 문장은 [10]에서 정의한 태그집합과 구문구조를 사용하여 수작업으로 구축한 말뭉치로서, 소설에서 추출한 853문장과 신문에서 추출한 544문장으로 총 1,397문장이고 15,148어절로 구성되어 있다. 그리고, 전체 문장 중의 90%는 학습집합으로 사용하고 나머지 10%를 실험집합으로 사용하였다. 각각에 대한 문장 수는 표 1과 같다.

표 1 실험집합

	학습집합	실험집합	총합
말뭉치	1,256문장	141문장	1,397문장

실험은 본 논문에서 제안한 신뢰도 기반의 구문패턴의 추출 및 적용하는 것이 얼마나 효과적인지를 알아보는 목적으로 수행되었다. 이 실험을 수행하기 위해서 표 2와 같은 네 가지 자질 집합을 설정하였다.

표 2 자질집합

자질집합	자질 조합
1	좌우 범주
2	좌우 범주+좌우 품사열
3	좌우 범주+좌우 품사열+좌우 어절수
4	좌우 범주+좌우 품사열+어휘열

실험에 대한 정확한 평가를 하기 위해서는 적용된 구문패턴이 올바른 것인지 틀린 것인지 확인하는 작업까지 고려하여야 하지만, 이는 수치적으로 측정하기 힘들기 때문에 고려대상에서 제외하였다. 이 실험에서 사용한 평가방법은 패턴 정확율, 패턴 재현율, 수작업 감소율이다. 패턴 정확율은 적용된 구문패턴들에 대해 정답의 비율을 나타낸다. 패턴 재현율은 정답구문구조에 대해 구문패턴의 올바른 적용 비율을 나타낸다. 수작업 감소율은 기존의 수동구축도구의 수작업 수에 대해 구문패턴이 수작업을 감소시킨 비율을 타나낸다. 이때, 구문패턴이 틀리게 적용되면 다시 수작업으로 구축할 뿐만 아니라 취소하는 수작업도 부가적으로 필요하므로 이를 고려하여 수작업 감소율을 계산한다. 이 세 가지 평가방법에 대한 수식은 다음과 같다.

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

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

$$\text{수작업감소율} = \frac{\text{맞은적용수} - \text{틀린적용수}}{\text{수동 구축도구에서의 수작업수}}$$

표 3 구문패턴 추출 및 적용결과

(수동구축도구에서의 수작업 수(a) = 2,658회)

항목 / 자질,신뢰도		추출된 패턴수	맞은 적용수 b	틀린 적용수 c	패턴 정확율 $\frac{b}{b+c}$	패턴 재현율 $\frac{b}{a}$	수작업 감소율 $\frac{b-c}{a}$
자질집합 1	50%	397개	1408회	6619회	17.5%	52.9%	-196.0%
	70%	273개	1414회	6130회	18.7%	53.1%	-177.4%
	90%	184개	1603회	3198회	33.3%	60.3%	-60.0%
	95%	162개	1595회	3099회	33.9%	60.0%	-56.5%
자질집합 2	50%	7739개	1555회	1324회	54.0%	58.5%	8.69%
	70%	2180개	1383회	596회	69.8%	52.0%	29.6%
	90%	871개	1142회	363회	75.8%	42.9%	29.3%
	95%	543개	1062회	291회	78.4%	39.9%	29.0%
자질집합 3	50%	14844개	1050회	235회	81.7%	39.5%	30.6%
	70%	2169개	773회	80회	90.6%	29.0%	26.0%
	90%	544개	595회	46회	92.8%	22.3%	20.6%
	95%	462개	532회	36회	93.6%	20.0%	18.6%
자질집합 4	50%	22500개	203회	4회	98.0%	7.6%	7.4%
	70%	728개	91회	0회	100.0%	3.4%	3.4%
	90%	92개	69회	0회	100.0%	2.5%	2.5%
	95%	54개	66회	0회	100.0%	2.4%	2.4%

실험은 각 자질집합에 대해서 50%, 70%, 90%, 95%의 신뢰도를 가지는 구문패턴을 선별하여 패턴 정확율, 패턴 재현율, 수작업 감소율을 조사하였다. 이 실험의 결과는 표 3과 같다.

4.2 실험결과

실험결과를 살펴보면, 본 논문에서 제안한 방법이 소량의 말뭉치에서 구문패턴을 추출하였음에도 불구하고 구문분석 말뭉치를 구축하는데 수작업을 30%정도 감소시킬 수 있다는 것을 알 수 있다.

각 자질집합 별로 실험결과를 분석하면 다음과 같다. 자질집합1은 좌우의 구문범주와 기능범주만을 고려하였기 때문에, 구문패턴이 잘못 적용된 경우가 많이 나타났다. 따라서, 구문패턴이 잘못 적용된 것을 취소하고 다시 분석하는데 오히려 더 많은 수작업이 요구되었다. 자질집합2는 자질집합1보다 사용한 정보가 더 많아서 패턴 정확율이 좋아지고 수작업도 다소 감소되었다. 자질집합3은 어절수까지 고려하므로 패턴 재현율은 자질집합2보다 떨어졌지만, 패턴 정확율은 크게 향상되어서, 약 30% 정도의 수작업 감소율이 나타내었다. 자질집합4는 어휘정보까지 고려하여 정확한 구문패턴들만이 추출되었지만 자료부족 문제가 심각하게 나타났다. 따라서 패턴 정확율이 매우 높음에도 불구하고 패턴 재현율이 너무 낮아서 수작업 감소율도 매우 낮았다.

신뢰도 별로 실험결과를 분석하면, 신뢰도가 높아질수록 패턴 정확율이 증가하고, 패턴 재현율이 감소한다는 것을 알 수 있다. 그리고, 수작업 감소율은 패턴 정확율 뿐만 아니라 패턴 재현율이 함께 영향을 끼친다.

위의 실험은 소량의 말뭉치에 대해서 학습하고 100문장에 대해서 실험하였기 때문에 정확한 구문패턴의 성능을 나타내지 못하였다. 더 많은 실험집합을 사용할 때, 수작업 감소율이 증가할 수 있는지를 알아보기 위해서 실험집합을 700문장부터 1200문장까지 100 문장 단위로 쪼개서 수작업 감소율을 실험하여 보았다. 실험은 수작업 감소율이 가장 높게 나타났던 자질집합3과 신뢰도 50%에 대해서만 수행하였다. 실험 결과는 그림6과 같다.

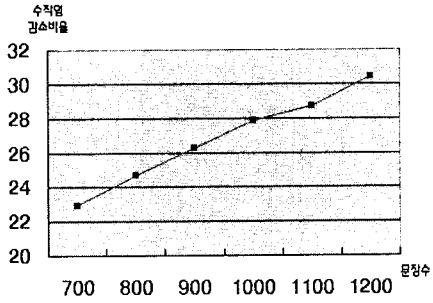

그림 6. 수작업 감소율의 변화

실험결과를 보면 알 수 있듯이, 실험집합의 양이 증가할수록 수작업 감소율도 더 증가할 것이라는 것을 알 수 있다. 앞으로 대량의 말뭉치가 구축된다면, 수작업 감소율도 더 개선되리라 기대된다.

5. 결론 및 향후 연구

본 논문에서는 구문패턴을 사용한 반자동 구문분석 말뭉치 구축도구를 제안하였다. 제안하는 방법은 구문범주, 품사열, 어휘열과 같은 자질들을 사용하여 구문패턴 후보를 추출하고, 이를 신뢰도를 사용하여 검증하였다. 이 방법은 다음과 같은 특징을 가진다.

첫째, 제안하는 반자동 구문분석 말뭉치 구축도구는 수작업을 감소시킬 수 있음을 실험을 통하여 증명하였다. 실험결과 이 방법은 약 30%정도의 수작업 감소율을 보였고, 실험집합의 양이 많아지면 더 많은 수작업 감소율을 얻을 수 있을 것이다.

둘째, 제안하는 방법은 기존의 구문분석 말뭉치를 사용하여 구문패턴을 자동으로 추출할 수 있다. 본 논문에서는 구문패턴을 추출하는 방법만을 제안하고, 실제 내용이 되는 구문패턴들은 기존에 구축된 말뭉치를 사용하여 추출하였다.

셋째, 제안하는 방법은 구문패턴을 이루는 자질집합을 선택하여 사용할 수 있다. 즉, 구

축되어 있는 말뭉치의 양이 많다면, 어휘와 같은 자세한 정보를 제공하는 자질을 사용하여 정확율을 높일 수도 있다. 반면에, 구축되어 있는 말뭉치의 양이 적다면, 품사열과 같은 일반적인 정보를 제공하는 자질을 사용하여 적용범위를 높일 수도 있다.

향후 작업으로, 더 적은 구문패턴을 가지고 더 넓은 적용범위를 가질 수 있는 구문패턴 일반화 방법을 연구하고, 더 정확한 구문분석 말뭉치를 구축할 수 있도록 자질집합, 신뢰도, 수작업 수, 구문패턴의 정확율 사이의 관계를 연구할 계획이다.

참고문헌

[1] 류원호, 이상주, 임해창, "어휘 문맥 의존 규칙과 통계 모델을 이용한 한국어 품사 부착 말뭉치 구축 도구", 정보과학회 논문지, 제25권, 1호, pp.396-398, 1998.

[2] 박소영, 곽용재, 정후중, 황영숙, 임해창, "한국어 구문분석의 효율성을 개선하기 위한 구문제약규칙의 학습", 정보과학회 논문지 (게재예정).

[3] Charniak, Eugene, "Tree-bank grammars". AAAI/IAAI Vol. 2, pp.1031-1036, 1996

[4] Michael Collins, "Head-Driven Statistical Models for Natural Language Parsing", PhD Dissertation, University of Pennsylvania, 1999.

[5] 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

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

[7] 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.

[8] 곽용재, 황영숙, 박소영, 정후중, 임해창, "결정적 구문 분석을 위한 문맥 의존 문법 규칙 획득 도구", 정보과학회 논문지, 제26권, 1호, pp.342-344, 1999.

[9] Hindle, Donald, "Acquiring disambiguation rules from text," in Proceeding. ACL, pp. 118-125, 1989.

[10] 김홍규 외, "제8장 구문 분석 방법론 및 표지의 권장 표준안 연구", 21세기 세종계획 국어 기초자료 구축 학술용역 과제 보고서, pp.377-403, 2001.

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