기계학습 기법을 이용한 한국어 구문분석

이용훈○ 이종혁
포항공과대학교 전자컴퓨터공학부 컴퓨터공학과
{yhlee95, jhlee}@postech.ac.kr

Korean Parsing using Machine Learning Techniques

Yong-Hun Lee○ Jong-Hyeok Lee
Department of Computer Science and Engineering
Division of Electrical and Computer Engineering
Pohang University of Science and Technology

요 약

최근의 구문분석 연구는 컴퓨터 성능 향상과 사용 가능한 대량의 구문분석 말뭉치 증가, 견고한 기계학습 기법 개발 등에 힘입어 통계적인 모델 연구가 꾸준히 증가하고 있다. 본 논문에서는 기존에 개발된 다양한 기계학습 기법 중 ME(Maximum Entropy) 모델과 SVM(Support vector machine) 모델을 이용한 한국어 구문분석 방법을 제안한다. 국어정보베이스(KIBS) 구문분석 말뭉치를 가지고 실험한 결과 SVM 모델을 이용한 한국어 구문분석기가 기존의 확률 기반 통계적 한국어 구문분석기의 성능보다도 최대 1.84% 높은 87.46%의 의존관계 결정 정확률을 보였다. 추후 언어지식을 반영한 다양한 자질들을 이용할 경우 성능 향상이 기대된다.

1. 서 론

구문분석은 문장의 구조를 파악하는 과정으로 기계번역, 정보추출, 질의응답시스템 등의 자연어처리 응용분야에서 핵심적 역할을 담당한다. 하지만 프로그래밍 언어와 같은 인공언어와는 달리 자연언어는 많은 애매성이 존재하여 문장의 정확한 구조를 파악하는 것이 쉽지 않다. 기존의 많은 연구들은 이러한 구조적인 중의성을 해결하기 위하여 규칙이나 확률 또는 기계학습 기법 등 다양한 방법을 이용하여 가장 올바른 문장의 구조를 찾으려는 노력을 기울여 왔다.

통계적인 방법론에 기반한 구문분석은 컴퓨터의 성능이 향상되고 대량의 말뭉치가 구축된 이후부터 꾸준히 개발되기 시작하였다. 결정트리(Decision tree), ME(Maximum Entropy), SVM(Support Vector Machine), CRF(Conditional Random Field) 등 기계학습 기법이 개발되고 널리 알려지게 된 이후에는 이들 기계학습 기법을 이용한 연구들이 활발히 진행되고 있다.

기존 한국어 통계기반 구문분석기들은 주로 대량의 구문분석 말뭉치로부터 확률을 학습하고 이를 이용하여 최적의 구문분석 결과를 선택하는 과정으로 이루어졌다.영어나 일본어 연구에는 이러한 순수한 확률 모델 외에도 기계학습 기법에 기반한 연구들이 많다. 본 논문의 모델이 되는 연구는 ME 모델에 기반한 일본어 의존구조 분석[1]과 SVM 모델을 이용한 통계적 의존관계 분석[2, 5] 두 가지이다.

Uchimoto는 일본어 분절의 다양한 자질들을 이용하여 각 자질들의 가중치를 구하고 이를 교토대학교 말뭉치에 적용한 결과 87.2%의 높은 의존관계 정확도를 얻었다고 한다[1]. ME 프레임워크에서는 주어진 자질들로 이루어진 여러 확률로부터 불확실성이 가장 큰 쪽 즉 엔트로피가 가장 커지도록 가중치를 학습하게 되는데 이는 인간의 사고에서 확실한 것 외에는 모두 같은 확률을 부여하는 것이 가장 공평하다는 원리와 같다. 실제로 ME는 형태소분석 및 태깅, 구문분석 등의 여러 NLP 분야에 사용되어 높은 성능을 보인 바 있다. 구문분석 알고리즘은 일본어의 특성 중 머리어 후위(head-final)[1] 원칙을 반영한 Sekine의 Backward beam search방법[3]을 이용하였다. Yamada는 영어를 대상으로 SVM 모델을 이용한 결정적 의존관계 구문분석기를 개발하였는데, SVM 모델을 이용하여 문장을 구성하는 각각의 단어들에 대해 양쪽 문맥의 다양한 자질들을 이용하여 Shift, Right, Left 세 가지 action을 결정하는 식으로 구문분석을 수행한다. 의존관계의 유무를 결정할 두 개의 단어에 대해 shift는 의존관계 판단을 보류하고 대상 단어를 오른쪽으로 이동하라는 것이며Right는 대상 단어들 중에 왼쪽 단어가 오른쪽 단어에 의존한다는 것을, Left는 반대로 오른쪽 단어가 왼쪽 단어에 의존한다는 것을 의미한다. 이 방법론을 이용하여 Penn Treebank에 적용한 결과 90.3%의 높은 의존관계 결정 정확률을 기록하였다[5]. 이는 기존에 개발된 최고 성능의 구문분석기와 비교해도 뒤지지 않을 만한 성능이다.

본 논문의 구성은 다음과 같다. 먼저 2절에서는 본 논문에서 제안하는 통계적 한국어 구문분석 방법에 대해 설명하고, 3절에서는 실험 및 분석결과를, 마지막으로 4절에서는 결론을 맺는다.

2. 기계학습 기법을 이용한 한국어 구문분석

본 논문에서 사용한 통계적 한국어 구문분석은ME 모델을 이용한 구문분석과 SVM 모델을 이용한 구문분석이다. 이들 구문분석은 한국어와 같이 부분적 자유어순을 가진 언어에 적합하다고 알려진 의존문법에 사용한다.각각의 기계학습 모델과 구문분석 알고리즘을 설명하기에 앞서 먼저 한국어 의존문법의 특징을 살펴보고 이들 특징이 가장 잘

1) 의존관계 중 지배소가 항상 뒤쪽에 나타난다는 원칙으로 한국어도 똑같은 특징을 가지고 있다.

반영된 최적의 문맥자질 집합(context feature set)과 한국어 구문분석 알고리즘을 설명하도록 한다.

2.1. 한국어 의존문법의 특징

한국어의 문장은 그 기본단위가 어절이다. 어절은 보통 하나의 내용어(content word)와 여러 개의 기능어(function word)가 접합되는데 이러한 특징 때문에 한국어는 교착어(agglutinative language)로 분류된다. 잘 발달된 조사나 어미와 같은 기능어를 통해 문장의 시제나 문장성분, 어절과 어절 간의 관계 등이 나타나기 때문에 한국어는 비교적 자유로운 어순을 가질 수 있으며 이 때문에 영어나 불어와 같은 언어와는 달리 주로 의존문법을 사용하여 구문분석을 하게 된다. 의존구문분석의 특징 중 가장 보편적인 특징에는 각 어절의 머리어(head)가 유일해야 한다는 것(uniqueness)과 각각의 의존관계는 서로 교차하지 않는다는 특징(projectivity)이 있다. 또한 한국어의 고유한 특징으로는 머리어 후위 원칙이 있다. 이에 따라 모든 어절의 머리어는 자신의 오른쪽에 존재하므로 구문분석 시 문장의 마지막 어절부터 앞쪽으로 구문분석을 진행함으로써 잘못된 구문분석 후보를 초기에 제거할 수 있게 된다. 각 어절 간의 의존관계 결정 시에 가장 유용하게 사용되는 자질은 기능어이다. 앞서 설명하였듯이 기능어는 그 자체만으로도 문장기능이나 어절 간의 관계를 나타내므로 문맥자질 집합을 선정할 때 이를 잘 반영함으로써 높은 구문분석 성능을 기대할 수 있을 것이다.

2.2. ME 모델을 이용한 구문분석

학습데이터가 주어질 경우, 이 학습데이터에 가장 적합한 모델 P를 찾는 것이 통계적 모델링의 주목적이라 한다면 ME 모델에서는 문맥정보 x가 주어졌을 때 결과 y가 나올 조건부 확률은 (식 1)과 같다[9].

$$p(y|x) = \frac{1}{Z(x)} exp\left[\sum_{i=1}^{k} \lambda_i f_i(x,y)\right] \qquad \text{(식 1)}$$

이는 조건부 확률을 파라미터 형태로 표현한 것으로 $f_i(x, y)$는 문맥정보를 나타내는 자질 함수(feature function)이며 λ_i는 이들 자질 함수들의 가중치(weight)를 나타낸다. (자세한 자질집합은 3.4절 참조)

주어진 자질집합과 학습데이터로부터 모델을 학습하기 위해서는 모델의 Log-likelihood를 최대로 하는 파라미터를 찾아야 한다. GIS(Generalized Iterative Scaling), IIS(Improved Iterative Scaling) 등의 학습 알고리즘이 있지만 최근에는 가장 최적화된 알고리즘으로 알려진 L-BFGS(Limited-Memory Variable Metric)을 주로 사용한다. 본 연구에서는 ME 알고리즘을 따로 구현하지 않고 Zhang Le의 C++버전 ME toolkit[9]을 사용하였다.

ME 모델을 이용한 구문분석은 한국어의 머리어 후위 원칙과 N개의 구문분석 결과를 출력할 수 있는 Backward beam search 방법을 이용하였다[3]. (그림 1)은 "그는 다시 파이를 만들어 그녀에게 주었다"라는 문장을 구문분석한 과정을 보인 것으로 Beam의 크기는 3으로 잡은 경우이다.

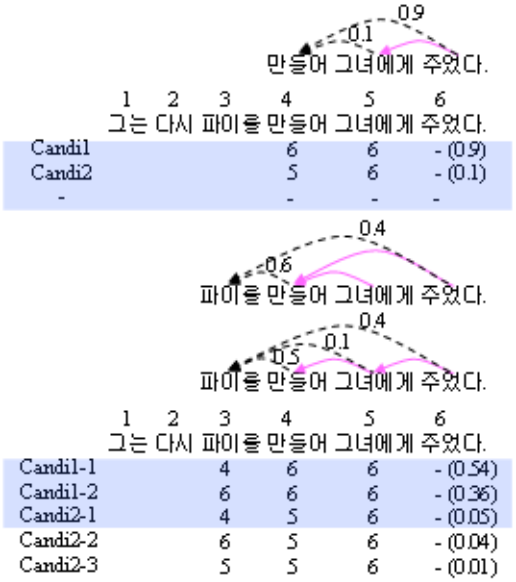

그림 1. Backward Beam Search 예제

구문분석은 문장의 마지막 어절부터 앞쪽으로 진행된다. 먼저 "주었다"어절은 문장의 머리어(head, root node)에 해당하므로 지나친다. 그 앞어절(n-1) "그녀에게"의 경우 머리어 후위 원칙에 따라 "주었다" 어절이 머리어가 된다. 다음 "만들어" 어절의 경우 ME 모델에 따라 "그녀에게" 어절이 머리어가 될 확률은 0.1이고 "주었다" 어절이 머리어가 될 확률은 0.9이다[2]. 이 두가지 경우는 각각 구문분석 결과 후보로서 확률에 따라 정렬하여 Beam 2개를 채우게 된다(네모 상자). 각각의 Beam에 대해서 앞의 과정을 반복하면 총 6개의 구문분석 결과 후보가 가능하지만 각각의 의존관계가 서로 교차하지 않는다는 의존문법의 특성을 적용하게 되면 5개의 후보만이 남게 된다. 이를 다시 확률에 따라 정렬하여 미리 설정한 Beam의 크기만큼의 결과만을 남기고 앞의 과정을 반복하게 되면 최종적으로 Beam의 크기에 해당하는 구문분석 결과를 얻게 된다. 만약 최적의 구문분석 결과만을 원한다면 확률이 제일 큰 것을 출력하면 된다.

2.3. SVM 모델을 이용한 구문분석

SVM 모델은 이진분류기의 하나로서 자연언어처리 분야에서 매우 폭넓게 사용되는 기계학습 기법 중 하나이다.

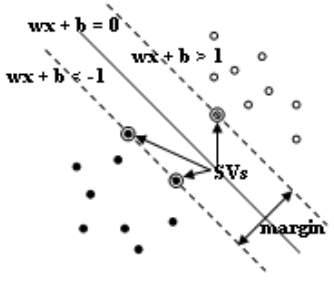

그림 2. SVM 모델 도식

2) 이 확률은 말뭉치의 실제 확률이 아닌 예제를 위한 가상 확률이다.

SVM 모델의 학습과정은 (그림 2)에서 보는 것과 같이 L개의 학습데이터 (x_i, y_i), ($1 \le i \le L$, x_i는 n차원의 자질벡터{feature vector}, $y_i \in \{+1, -1\}$)가 주어졌을 때, 이 학습데이터를 분할하는 초평면(hyperplane) $w \cdot x + b = 0$ 중에서 최대의 마진(maximum margin)을 가지는 초평면을 찾는 과정이며, 최적의 초평면은 (식 2) 조건을 만족하는 quadratic programming 문제를 풀음으로써 찾을 수 있다. (C는 상수, ξ_i는 선형분리가 불가능한 데이터를 위한 slack variable)

$$\min \frac{1}{2} \| w \| + C \sum_{i=1}^{l} \xi_i$$
$$s.t.\ y_i(w x_i + b) \ge 1 - \xi_i,\ \xi_i \ge 0 \qquad \text{(식 2)}$$

이렇게 구한 초평면 f(x)는 자질벡터를 고차원의 벡터공간으로 사상시켜 선형분리가 가능하도록 하는 커널함수(K)를 이용한 일반화된 형태의 부호함수인 (식 3)으로 표현할 수 있다[5].

$$f(x) = sign\left(\sum_{i: x_i \in SVs} \alpha_i y_i K(x_i, x) + b \right) \qquad \text{(식 3)}$$

앞서 설명하였듯이 한국어는 머리어 후위 원칙을 가지고 있다. 이러한 한국어의 언어적 특징을 고려할 경우 구문분석의 action은 shift, depend 두 가지로 정의할 수 있으며 이진분류기인 SVM 모델을 그대로 사용할 수 있게 된다. (그림 3)은 앞서 ME 모델에서 사용한 문장 중 "파이를 만들어 그녀에게 주었다" 부분을 SVM 모델을 이용하여 구문분석하는 과정을 그림으로 나타낸 것이다. "파이를" 어절과 "만들어" 어절은 SVM 모델의 예측 결과 Depend action이 결정된다. 따라서 "파이를" 어절은 "만들어" 어절의 의존소가 된다. 다음으로 대상이 되는 어절은 "만들어"와 "그녀에게" 어절이며 SVM 모델은 Shift action을 예측하게 된다. 이는 두 어절 간에 의존관계가 존재하지 않기 때문이다. 이러한 과정을 문장의 머리어(head, root node)인 마지막 어절을 제외한 모든 어절이 머리어를 가질 때까지 반복한다.

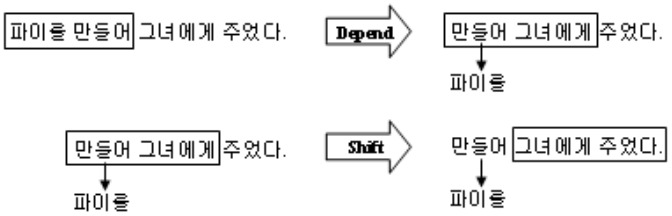

그림 3. SVM 모델을 이용한 구문분석

2.4. 문맥자질 집합(context feature set)

본 논문에서 사용하는 문맥자질은 가장 기본적인 어절 자체의 정보로 어절의 어휘(surface form)와 어절을 구성하는 전체 형태소들의 어휘와 품사(POS tag), 어절의 타입(내용어 및 기능어 타입), 내용어 및 기능어들의 어휘와 품사(한개 이상의 형태소로 구성될 수 있다)를 사용하였으며, 어절과 어절 간의 자질로서 어절 간의 거리, 어절 사이에 존재하는 괄호 정보, 어절 사이에 대상이 되는 좌/우 어절과 똑같은 형태의 어절이 존재하는지의 여부를 사용하였다. 또한 이 두 가지 자질 중에 2개에서 4개의 자질을 조합하여 만든 자질을 사용하였다. (표 1)은 본 논문에서 사용한 문맥자질들을 정리한 것이다.

표 1. 문맥자질 집합(context feature set)

구분	번호	문맥자질	설 명
어절 자체 정보 – 앞어절 자질: n, 뒤어절 자질: n'	1	어절 어휘	
	2	어절 타입	체언, 용언, 관형어 등
	3	형태소 어휘열	
	4	형태소 품사열	
	5	내용어 어휘열	
	6	내용어 품사열	
	7	기능어 어휘열	
	8	기능어 품사열	
	9	기호 종류	none ' " , . ! ? 등
	10	여는 괄호	({ [< 등
	11	닫는 괄호	) }] > 등
어절 간 정보	12	어절간의 거리	
	13	괄호 정보	어절 사이에 존재하는 괄호
	14	쉼표 유무	0, 1
	15	보조사 유무	0, 1
	16	앞어절과 동일한 어절 존재 여부	0, 1
	17	뒤어절과 동일한 어절 존재 여부	0, 1
조합 정보	18	8 + 2'	
	19	2 + 2'	
	20	8 +16 + 2'	
	21	8 +17 + 2'	
	22	10 + 13 + 11'	
	23	8 + 16 + 17 + 2'	
	24	8 + 9 + 8' + 9'	

이들 자질은 한국어의 언어적 특성이 반영된 자질들이다. 예를 들면, 각 어절을 내용어와 기능어 부분(5~8)으로 따로 분류하여 자질을 만든 것이나, 15번과 같이 잘 발달된 조사정보를 문맥자질로 사용한 것은 이들 정보가 두 어절 간의 의존관계를 파악하는데 있어 매우 중요하게 작용하기 때문이다.

3. 실험결과 및 분석

2절에서 제시한 두 가지 구문분석 모델은 국어정보베이스(Korean Language Information Base) 구문분석 말뭉치를 사용하여 학습하고 평가하였다. 총 12,084 문장 중 랜덤하게 뽑은 10,876문장(90%)을 학습데이터로 사용하여 모델을 구축하였고 나머지1,208 문장(10%)을 평가데이터로 사용하였다. (표 2)는 ME 모델을 사용한 통계 구문분석기의 성능을 Beam의 크기를 변화해 가면서 평가한 결과이다. 표의 결과에서 알 수 있듯이 Beam의 크기가 1일 경우 가장 높은 성능이 나타났다. 이는 한국어의 경우,

표 2. Beam 크기에 따른 ME 모델 구문분석 성능

Beam 크기	의존관계 정확률(%)
1	80.53
2	80.48
3	80.06
4	79.90
5	79.90
10	79.90

두 어절 간의 의존관계가 뒤쪽에서 차례차례 결정적으로 판단하여 구문분석 결과를 구하는 경우나 전체 문장을 모두 고려해서 가장 확률이 높은 것을 구하는 것에 차이가 없다는 것을 의미한다. 또한 한국어는 기능어 때문에 좌우 문맥정보 중 오른쪽 문맥정보가 더 중요하게 작용한다는 것을 간접적으로 보여주는 결과라 할 수 있다. 두 번째로 SVM 모델을 이용한 구문분석 성능과 기존에 통계적인 방법을 사용한 한국어 구문분석기 성능을 비교한 결과는(표 3)과 같다.

표 3. 구문분석 성능 비교

구분	의존관계 정확률(%)	비고
SVM(제안방법)	**87.46**	
ME(제안방법)	**80.53**	
SVM(Kudo)	89.09	일본어
SVM(Yamada)	90.30	영어
ME(Uchimoto)	87.20	일본어
Lex. Parser(Chung)	86.74	한국어
UnLex. Parser(Chung)	85.62	한국어

본 연구에서 제안한 SVM 모델을 이용한 한국어 구문분석 성능은 87.46% 의존관계 정확률을 보였다. 이는 비슷한 방법을 사용하여 일본어와 영어에 적용한 Kudo와 Yamada의 구문분석기의 성능보다 낮은 성능을 보였다. 이는 한국어의 어절이 일본어와 영어보다 복잡하게 구성을 가지며 중의성이 높아서 얻어진 결과이다. 실제로 구문분석 오류의 많은 부분이 보조사를 사용한 어절의 잘못된 머리어 선택[3]이라든가, 대등접속구문을 따로 처리하지 않음으로써 잘못된 머리어를 선택하는 경우에서 빚어진 것을 보면 이에 대한 추가적인 성능 보완이 필요하다. 한국어의 특징과 중의성을 해결할 수 있는 문맥자질을 보완한다면 다른 언어와 같은 높은 성능을 얻을 수 있으리라 기대한다.

하지만 똑같은 실험데이터를 사용한Chung의 통계적 한국어 구문분석기의 성능이 각각85.62%, 86.74%인 것을 감안하면 최대 1.84% 높은 성능으로서 제안방법이 매우 높은 성능을 보이고 있음을 알 수 있다[7]. 실험에 사용한 ME, SVM toolkit은 각각 [8], [9]이다.

3) 보통 '은/는'과 같은 보조사는 주제(topic)를 나타내기 위해 어절을 문장의 처음으로 보내 장거리 의존관계(long dependency)를 유발시킨다.

4. 결론

본 논문에서는 다양한 기계학습 기법 중에서ME 모델과 SVM 모델을 이용한 한국어 구문분석 방법을 제안하고 그 실험결과를 제시하였다. 실험을 통해 SVM 모델을 이용한 구문분석은 최대 87.46%의 의존관계 정확률을 보였다. 이는 비록 일본어나 영어에 비해 다소 낮은 성능이지만 한국어의 기본단위인 어절이 복잡하게 구성된다는 점과 문장의 구성성분인 주어나 목적어 등의 성분이 쉽게 생략된다는 점, 보조사와 같은 기능어 때문에 발생하는 중의성 등으로 인해 구문분석이 좀 더 어렵다는 점 때문에 기인한 것으로 보인다. 추후 추가 연구를 통해 이러한 문제점을 보완해 나갈 예정이다.

하지만 이러한 성능은 본 논문에서와 동일한 문서집합에서 구문분석을 해 높은 성능을 얻은Chung의 구문분석기와 비교해 최대 1.84% 높은 성능으로서 제안방법이 좀 더 우수함을 입증한다. 향후 본 논문에서 제안한 기계학습에 기반한 한국어 구문분석 방법과 기존의 사전과 규칙에 기반한 구문분석 방법론을 하나로 합친 하이브리드 방법론을 연구할 예정이다.

감사의 글

본 논문은 2008년도 두뇌한국21사업의 지원을 받았고 지식경제부 및 정보통신진흥연구원의 정보통신선도기반기술개발사업의 연구결과로 수행되었습니다.

참고 문헌

[1] Uchimoto, K., Sekine, S., Isahara, H., "Japanese dependency structure analysis based on maximum entropy models", In Proceedings of the European Association for Computational Linguistics, pp 196-203, 1999.

[2] Kudo, T., Matsumoto, Y., "Japanese dependency structure analysis based on support vector machines", In Empirical Methods in Natural Language Processing and Very Large Corpora, pp 18-25, 2000.

[3] Sekine, S., Uchimoto, K., Isahara, H., "Backward beam search algorithm for dependency analysis of Japanese", Proceedings of the 18th conference on computational linguistics - vol. 2, pp 754-760, 2000.

[4] Singhal, Amit, "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4), pp 35-43, 2001.

[5] Yamada, H., Matsumoto, Y., "Statistical dependency analysis with support vector machines", In Proceedings of IPWT, 2003.

[6] Matsumura, A., Takasu, A., Adachi, J., "Effect of relationships between words on Japanese information retreival", ACM Transactions on Asian Language Information Processing (TALIP), vol. 5, issue 3, pp 264-289, 2006.

[7] Hoojung Chung, "Statistical Korean Dependency Parsing Model based on the Surface Contextual Information", Ph.D. dissertation, 2004.

[8] Thorsten Joachims, Support Vector Machine Toolkit: SVMlight http://svmlight.joachims.org/

[9] A simple C++ library for maximum entropy classification, http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/maxent/