SVM을 이용한 절-절 간의 의존관계 설정

김 미 영†

요 약

문장이 길어질수록 구문분석의 정확률이 급격히 떨어지므로, 문장을 분할하여 각각의 분할단위로 구문분석을 수행한 후 각 구문분석결과를 합쳐 완성된 구문트리를 만드는 것이 일반적이다. 이 때 주로 절 단위로 문장이 분할되고, 각 절의 구문분석결과를 통합하게 되는데, 통합 과정에서 절-절 간의 의존관계 설정에 많은 오류가 생긴다. 이러한 절 간의 의존관계의 애매성을 해결하기 위하여, 본 논문은 기계학습을 이용하여 절-절 간의 의존관계를 설정하고자 한다. 따라서 이에 필요한 자질들이 무엇인지 알아보고, 성능향상에 기여를 하는 자질과, 오히려 성능을 저하시키는 자질들을 분석해 본다. Support Vector Machines(SVM)을 사용하여 성능을 평가하고, 본 논문에서 실험한 방법과 기존의 방법들의 성능을 비교해 본 결과, 절-절 간의 의존관계 설정에 있어서 8.88~15.35%의 성능향상을 보였다.

키워드 : 절, 의존관계, 구문트리, 구문분석, SVM

Determining the Dependency among Clauses based on SVM

Mi-Young Kim†

ABSTRACT

The longer the input sentences, the worse the syntactic parsing results. Therefore, a long sentence is first divided into several clauses, and syntactic analysis for each clause is performed. Finally, all the analysis results are merged into one. In the merging process, it is difficult to determine the dependency among clauses. To handle such syntactic ambiguity among clauses, this paper proposes an SVM-based clause-dependency determination method. We extract various features from clauses, and analyze the effect of each feature on the performance. We also compare the performance of our proposed method with those of previous methods.

Key Words : Clause, Dependency Relation, Syntactic Analysis, SVM

1. 서 론

입력 문장이 길수록 구문분석의 애매성이 증가하기 때문에 구문분석 성능은 급격히 떨어진다. 따라서 장문을 대상으로 구문분석을 할 경우 문장을 분할하여 작은 단위로 나눈 후, 각 단위별로 구문분석을 수행한 결과를 통합하여 하나의 완성된 분석결과를 내는 것이 일반적이다. 이 때 장문을 나누는 단위가 주로 절(clause)이고[1, 2, 3, 4, 5], 절 단위로 문장을 분할한 후 각 절의 구문분석을 우선 수행한다. 그 후 각 분석결과를 통합하여 하나의 구문분석 결과를 얻게 된다. 이러한 통합과정에 있어서 절-절 간의 의존관계 설정이 이루어져야 하는데, 이 때 많은 중의성이 존재한다. 따라서 본 논문은 기계학습을 이용하여 절-절 간의 의존관계를 설정하고자 하고, 이 때 필요한 자질들 및 각 자질이 성능에 영향을 미치는 정도를 각각 분석하고자 한다. 마지막으로, 본 논문에서 제안한 방법과 기존의 방법들과의 성능을 비교함으로써 본 논문의 제안방법이 효과적임을 보인다.

이 논문은 다음과 같이 구성되어 있다. 2장에서 기존연구 설명이 있을 것이고, 3장에서는 본 논문에서 제안하는 Support Vector Machines(SVM) 학습을 이용한 절-절 간의 의존관계 설정방법을 설명하고, 이에 필요한 자질을 설명할 것이다. 4장에서 본 논문에서 제안하는 방법을 이용한 실험결과 및 실제 성능향상에 영향을 미치는 좋은 자질들 및 오히려 성능을 저하시키는 자질들에 관한 분석이 이루어지고, 기존의 방법들과의 성능비교를 해본다. 마지막으로 결론 및 향후연구가 이어진다.

2. 기존연구

의존문법을 이용한 구문분석과정에서 절-절 간의 의존관계 애매성은 심각한 문제 중의 하나이다. 그래서, 기존의 많은 학자들이 절의 포함관계를 인식함으로써 의존관계의 문

※ 이 논문은 2007년도 성신여자대학교 학술연구조성비 지원에 의하여 연구되었음.
† 정 회 원 : 성신여자대학교 컴퓨터정보학부 전임강사
논문접수 : 2006년 9월 18일, 심사완료 : 2006년 12월 13일

제점을 해결하고자 하였다.

Shirai[6]는 절의 범위를 판단하는 데 있어서 가장 큰 영향을 미치는 것은 절의 서술어의 마지막 기능어라고 보고, 기능어 54개를 추출한 후 이것을 다시 3개의 타입으로 분류하였다. 그리고 언어학적 지식에 의거하여 "타입A < 타입B < 타입C" 라는 타입별 포함관계를 정리하였다. 하지만 이 연구는 단순한 규칙에 의한 것으로서 적용범위가 제한적이라는 단점을 가지고 있다. 노마 히데키[9]는 종속절의 상호 포섭관계에 대한 연구결과로, 한국어 종속연결어미 176개의 타입을 크게 10개의 절 타입으로 분류하여, 각 절 간의 포함관계를 휴리스틱을 토대로 하여 표로 만들어 보여주었으나, 역시 많은 예외문장이 존재하는 실정이다. 그 후 Danlos[10]는 프랑스어를 대상으로 두 개의 절 사이의 관계로서 의존관계를 연구하였는데, "절1 접속어1 절2 접속어2 절3" 의 경우 두 가지 구문관계가 있을 수 있으나, 절2와 접속어2 사이에 쉼표가 등장하면 절1, 절2, 절3의 의존관계가 중의적이지 않다고 설명하였다. 따라서 의존관계를 설정하는 데 있어서 쉼표가 중요한 정보가 된다는 것을 시사하고 있다. Y.H.Roh[11]는 영어를 대상으로 구문분석하는 과정에서 절의 의존관계를 설정하는 과정을 설명하였고, 이 때 쉼표와 접속사정보에 의거한 규칙을 만들어 사용하였다. 이러한 규칙 기반 방법들은 적용률이 높지 못하고, 실제 많이 나타나는 예외사항들을 처리할 수 없는 단점이 있다.

규칙 기반 방법의 한계를 극복하기 위하여, Utsuro[12]는 결정리스트를 통한 기계학습을 통해 통계적인 방법으로 절의 의존관계를 결정하고자 하였다. 여기서 구두점, 문법태그, 절의 마지막 단어의 활용형, 어휘 등의 4가지를 자질로 사용하여 두 개의 절 사이의 포함관계를 학습하였다. 그 후 Kawahara[7] 또한 결정리스트를 사용한 통계적 방법을 시도하였고, 자질로 서술어의 마지막 어미의 표층형태와 쉼표정보를 사용한다. 기존의 Shirai[6]와 Minami[8]는 동사의 어미의 표층형태를 여러 개의 클래스로 나누어 분류 후 사용하였는데, Kawahara[7]는 표층형 그대로 사용한 것이 차이점이다. Kawahara[7]은 서술어들을 클래스로 분류하여 사용하지 않았고, 휴리스틱에 의해 서술어들간의 포함관계를 설정하지 않았기 때문에 작위적이지 않고 객관성을 유지한다. 하지만 위의 연구들에 있어서, 각 자질에 대한 중요도의 설명이 부족하다. 또한 방법 적용 후 절-절 간의 의존관계에 대한 기존의 다른 방법과의 성능비교가 없다. 따라서 본 논문에서는 절로부터 적절한 자질을 추출하여 Support Vector Machines(SVM) 방법을 이용하여 실험함으로써, 기존의 방법과의 성능을 비교해 보고, 제안한 자질들이 성능에 미치는 영향을 분석하여 실제 가장 좋은 자질이 무엇인지 알아본다.

3. 절-절 간의 의존관계 설정

3.1 SVM을 이용한 기계학습

절-절 간의 의존관계 설정을 위해 어떤 기계학습 방법이라도 적용가능하지만, 본 논문에서는 Support Vector Machines(SVM)을 이용하여 실험한다. SVM은 많은 자연언어처리 응용분야에 사용되어 좋은 결과를 보여주었다[13], [14,15]. SVM은 적은 데이터의 학습에 강한 특징을 가지고 있고, 학습데이터가 매우 큰 차수를 가지고 있어도 좋은 일반화 성능을 나타내고 있다. SVM은 2개의 클래스를 대상으로 하는 분류기이므로, 3개 이상의 지배소 후보절들 중 하나를 선택하는 절-절 간의 의존관계 설정에 사용하기 위해서는 여러 개의 클래스를 다룰 수 있도록 SVM을 확장해야 한다. 따라서 3개 이상의 클래스문제를 다루도록 만들어진 LIBSVM[16] 프로그램을 사용하여 실험하였다. 지금부터 지배소 절을 찾고자 하는 절을 '의존소 절'이라고 명명할 것이며, 의존소 절의 지배소가 되는 절후보로는 의존소 절의 오른쪽에 가까이 위치한 절들 중 6개의 절을 대상으로 한다.

3.2 절-절 간의 의존관계 설정을 위해 추출한 자질들

의존관계를 설정함에 있어서 위치정보가 아주 중요하다. 가까이 존재할수록 의존관계가 발생할 가능성이 크기 때문이다. 실제 의존소 절의 지배소가 되는 절의 위치는 가장 가까운 오른쪽인 경우가 많으나, 주변 절들과의 포섭관계에 의거하여 거리가 떨어진 절들이 지배소가 되는 경우 또한 많이 발생한다. 따라서 절의 오른쪽 가장 가까운 절부터 6번째로 떨어져 있는 절까지를 대상으로 하여, 한꺼번에 자질정보를 추출하여 서로간의 포섭관계를 고려한 후, 실제 정답 지배소 절의 위치를 학습하도록 한다. 클래스 값은 6개(1~6)로 구성되게 되고, 각 클래스값은 지배소로 결정된 절의 위치를 나타낸다. 절들로부터 자질 추출시 데이터 부족 문제를 염려하여 표층형태 뿐 아니라 의미클래스까지 자

〈표 1〉 학습에 사용된 자질들

첫 번째 자질	서술어 내용어의 품사
두 번째 자질	서술어 마지막 어미의 표층형태
세 번째 자질	서술어 내용어의 의미코드
네 번째 자질	의존소 절과 같은 주어의 공유 여부
다섯 번째 자질	쉼표정보

〈표 2〉 각 자질의 값들

자질 종류	자질값
첫 번째	보통명사, 고유명사, 의존명사, 인칭대명사, 재귀대명사, 양수사, 서수사, 일반동사, 지시동사, 성상형용사, 지시형용사, 그외
두 번째	ㄴ, ㅁ, 기, 음, ㄴ데, ㄴ즉, ㄴ지, ㄹ지, ㄹ지니, 거나, 게, 고, 나, 는데, 는지, 니, 다가, 도록, 든지, 듯이, 라, 려고, 며, 면, 면서, 므로, 아, 아서, 어, 어도, 어서, 어야, 으나, 으니, 으려고, 으며, 으면, 으면서, 으므로, 자, 지, 지마는,… (모든 연결어미)와 종결어미(ㅂ니다, ㅂ니까, ㄴ다 등등), 관형형어미 등
세 번째	카도가와 시소러스 의미코드
네 번째	1, 0
다섯 번째	1, 0

질에 추가하도록 한다. 각 자질은 아래에서 상세히 설명하도록 한다. 절들의 중심어는 서술어이므로 서술어의 정보에서 주로 자질이 추출된다. <표 1>에서 보여지는 바와 같이, 5가지 종류의 자질을 사용하게 된다. 첫 번째 자질은 서술어의 내용어의 품사정보이다. 실제 동사, 형용사 뿐 아니라 '체언+하다', '체언+이다'로 쓰여서 서술어로 사용되는 경우도 있으므로, <표 2>에서 보이고 있는 12개의 품사정보 중 하나가 할당된다. 두 번째 자질은 서술어의 어미의 표층형태를 이용한다. 한국어는 교착어이고, 서술어의 어미는 다음 절과의 연결적 기능을 나타내므로 절-절 간의 의존관계 설정에 있어서 중요한 자질이 될 수 있다. 세 번째 자질로서 서술어의 의미코드가 사용된다. 의미코드는 카도카와 시소러스의 표현방법을 따르고 있으며, 이는 1110개의 의미클래스가 트리구조로 할당되어 있다. 지금까지의 3개의 자질들이 서술어에서 추출한 것이라면, 4번째 자질은 절 자체에 관련된 자질이다. 이 자질은 지배소 후보 절이 의존소 절과 같은 주어를 공유하는지 나타낸다. 실제 하나의 주어가 여러 절에 공통되어 사용될 때, 문장에서 그 주어는 한 번만 등장하고 공유하여 쓰는 경우가 대부분이다. 실제로는 의존소 절의 오른쪽에 가장 가까이 위치하는 절이 지배소인 경우가 많으나, 같은 주어를 공유하는 절이 의존소 '과 떨어져서 문장에서 등장하는 경우에는 이 절이 의존소 절과 연관되어 의존관계를 맺을 수 있다. 따라서 지배소 절 후보가 주어를 공유하는지에 관한 정보를 자질로 사용한다. 각 절에서 주격조사와 '~은/는' 보조사가 기능어로 사용된 체언을 주어로 인식한다. 실제 '~은/는' 보조사가 붙은 체언이 주어 역할을 하는 경우가 다른 문장성분의 역할을 하는 경우보다 훨씬 더 많은 비중을 차지하므로, 모두 주어로 가정하여 실험하였다. 그리고 주어 기능을 하는 단어가 없는 절의 경우는 Leffa[17]에서와 같이 앞쪽 가장 가까운 절의 주어를 공유하는 것으로 가정하여, 그 절의 정보를 가져온다. 다섯 번째 자질은 절이 쉼표로 끝나는지 여부를 나타낸다. 문장에서 중간에 쉬기 위해 사용되는 쉼표는 절의 지배소를 파악하기 위한 유용한 특징이 된다. 이렇게 다섯 가지의 자질을 학습을 위해 사용하고, 각 자질에 대한 상세한 값들은 <표 2>에 나타나 있다.

우리는 7개의 절- 의존소 절과 의존소 절의 오른쪽에 인접한 6개의 절 -- 로부터 위의 5개의 자질들을 추출하여 사용한다. 따라서 클래스 값은 6개(1~6)로 구성되게 되고, 각 클래스값은 지배소로 결정된 절의 위치를 나타낸다.

4. 실 험

우리는 제안된 절-절 간의 의존관계 방법을 실험하기 위해 국어정보베이스[1] 말뭉치의 교양산문 관련 문장들을 대상으로 실험한다. 22,900개의 한국어 문장 (평균 14.08어절/문장)을 대상으로 10-교차확인법 (10-fold cross validation)으로 실험하였다.

1) http://kibs.kaist.ac.kr/

〈표 3〉 절-절 간의 의존관계 성능 비교

	본 논문의 SVM기반 방법	가장 가까운 절을 지배소로 결정하는 방법	노마히데키[9]의 포섭관계 규칙 사용
절-절 간의 의존관계 정확률	81.94%	66.59%	73.06%

〈표 4〉 각 자질을 사용하지 않았을 경우 성능변화

사용한 자질들	정확률 변화
5개의 자질 모두 사용	0.00%
품사정보 제외한 4개의 자질 사용	+1.01%
표층형태 제외한 4개의 자질 사용	-3.32 %
의미정보 제외한 4개의 자질 사용	+1.38%
주어공유정보 제외한 4개의 자질 사용	-1.96%
쉼표정보 제외한 4개의 자질 사용	-2.60%
5개의 자질 사용 (표층형태 대신, 표층형태를 5개의 클래스로 분류한 후 클래스정보 사용)	-2.57%

우리는 아래 3가지 성질을 분석했다.

1. 다섯 개의 자질들 중 하나를 사용하지 않았을 경우 성능변화
2. 어말어미의 표층형태를 그대로 사용하였을 경우 vs. 표층형태를 5개의 클래스(관형형, 대등형, 종속형, 종결형, 인용형)로 분류한 후 클래스 정보를 사용하였을 경우
3. 기존의 규칙기반 방법들 vs. 본 논문에서 제안한 방법

실험 결과, 우리는 아래의 결과들을 얻었다.

1. 가장 중요한 자질은 '어말어미의 표층형태'이고, 의미정보와 품사정보는 성능을 오히려 떨어뜨리는 나쁜 자질이다. (<표 4> 참조)
2. '어말어미의 표층형태' 자질 대신, 표층형태를 다섯 개의 클래스(관형형, 종결형, 대등형, 종속형, 인용형)로 분류한 후 분류정보를 사용하였을 경우 성능이 떨어진다. (<표 4> 참조)
3. SVM을 이용한 본 논문의 방법이 기존의 규칙기반 방법보다 성능이 좋았다. (<표 3> 참조)

<표 4>에서 알 수 있듯이, 의미정보와 내용어의 품사정보는 제외하였을 경우 오히려 성능향상을 가져왔다. 다시 말하면, 절-절 간의 의존관계 설정에 있어서 의미정보와 내용어의 품사정보는 중요하지 않다.

그 외 세 개의 자질 - 어말어미의 표층형태, 주어공유 정보, 쉼표정보 - 은 성능향상에 기여를 하는 자질들로 판명이 났고, 특히 어말어미의 표층형태를 제외하였을 경우 가장 크게 성능이 저하되었으므로, 어말어미의 표층형태 정보가 절-절 간의 의존관계에 있어서 가장 큰 영향을 주는 자질로 분석된다. 앞서 언급하였듯이, 어말어미의 표층형태 정보는 다음 절과의 연결기능을 나타낸다. (예를 들어 '~으므로'는 현재의 절이 다음 절의 '이유'로서 기능함을 나타낸다.) 따라

서, 이 정보가 절-절 간의 의존관계 설정에 있어 가장 큰 영향을 주었다고 추측할 수 있다.

<표 4>에서 알 수 있듯이, 표층형태 정보를 그대로 사용하지 않고, 5개의 클래스로 분류한 후 클래스정보를 대신 사용하면, 성능이 저하됨을 알 수 있다. 클래스 정보로 대치하여 사용시, 분류가 촘촘하지 않으므로 정보의 손실이 발생한다는 것을 유추할 수 있다.

본 논문의 방법을 기존의 방법들과 성능비교하기 위하여, 두 가지의 실험을 추가했다. 하나는, 가장 가까운 절을 지배소로 결정하는 방법(nearest modifiee principle)을 사용한 실험이고, 또 한 가지 방법은 노마 히데키[9]의 절-절 간의 포함관계 관련 규칙을 그대로 사용하였을 경우이다. <표 3>에서 볼 수 있듯이, SVM을 이용한 기계학습 방법이 기존의 방법들보다 좋은 성능을 보였다.

5. 결 론

이 논문은 절-절 간의 의존관계의 설정을 위해 SVM 을 이용한 기계학습방법을 제안하고, 이에 유용한 자질들이 무엇인지 분석한다. 본 논문에서의 방법은 기존의 규칙 기반 방법들보다 성능이 앞섰고, 어말어미의 표층형태, 쉼표, 그리고 주어공유 정보가 의존관계 설정에 좋은 자질임을 분석했다. 이 중 표층형태가 가장 큰 영향을 주는 자질이었고, 의미정보와 내용어의 품사정보는 성능을 오히려 감소시키는 자질임을 보였다.

앞으로 본 연구를 다음과 같이 계속해 나갈 계획이다. 우선, 위의 방법을 적용한 후 절-절 간의 의존관계 오류의 타입을 분석하여 성능을 더욱 향상시킬 방법을 연구해 본다. 두 번째로는 타 언어 또한 절-절 간의 의존관계에 있어 애매성을 가지므로 타 언어를 대상으로 위의 자질을 사용한 방법을 적용하여 성능을 분석해 볼 계획이다.

참 고 문 헌

[1] X. Carreras, L. Marquez, V. Punyakanok, and D. Roth, "Learning and inference for clause identification", Proc. 13th European Conference on Machine Learning, Helsinki, Finland, pp.35-47, 2002.

[2] V. J. Leffa, "Clause processing in complex sentences", Proc. 1st International Conference on Language Resources and Evaluation, Granada, Spain, 1998.

[3] A. Molina and F. Pla, "Clause detection using HMM", Proc. 5th Conference on Computational Natural Language Learning, Toulouse, France, pp.162-164, 2001.

[4] E. F. T. K. Sang and H. Dejean. "Introduction to the CoNLL-2001 shared task: clause identification". Proc. CoNLL-2001, Toulouse, France, pp. 53-57, 2001.

[5] X. Carreras and L. Marquez, "Boosting Trees for Clause Splitting", Proc. CoNLL-2001, Toulouse, France, pp.73-75, 2001.

[6] S. Shirai, S. Ikehara, A. Yokoo, and J. Kimura. "A new dependency analysis method based semantically embedded sentence structures and its performance on Japanese subordinate clauses." Transactions of Information Processing Society of Japan, 36(10):2353-2361, 1995.

[7] D. Kawahara and S. Kurohashi. Corpus-based Dependency Analysis of Japanese Sentences using Verb Bunsetsu Transitivity, In Proceedings of the 5th Natural Language Processing Pacific Rim Symposium, pp. 387-391, 1999.

[8] F. Minami. "Gendai Nihongo no Kouzou'(structures of Modern Japanese Language), Taishuukan shoten, 1974.

[9] 노마 히데키, "한국어 어휘와 문법의 상관구조", 태학사, 2002.

[10] L. Danlos, "Sentences with two subordinate clauses : syntax, semantics and underspecified semantic representation'. In Proceedings of the TAG+7 Workshop, p.140-147, 2004.

[11] Yoon-Hyung Roh, Young Ae Seo, Ki-Young Lee, Sung-Kwon Choi: Long Sentence Partitioning using Structure Analysis for Machine Translation. Proceeding on NLPRS, p.646-652, 2001.

[12] T. Utsuro, S. Nishiokayama, M. Fujio, and Y. Matsumoto, "Analyzing dependencies of Japanese subordinate clauses based on statistics of scope embedding preference," Proc. 1st Conference of the North. American Chapter of the ACL, pp.110 - 117, 2000.

[13] T. Kudo and Y. Matsumoto, "Chunking with Support Vector Machines", Proc. 2nd meeting of North American Chapter of Association for Computational Linguistics (NAACL), Pittsburgh, PA, USA, pp.192-199, 2001.

[14] T. Kudo and Y. Matsumoto, "Japanese Dependency Structure Analysis Based on Support Vector Machines", Proc. 2000 SIDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), Hongkong, China, pp.18-25, 2000.

[15] H. Yamada and Y. Matsumoto, "Statistical Dependency Analysis with Support Vector Machines", Proc. 8th International Workshop on Parsing Technology, Nancy, France, pp.195-206, 2003.

[16] Chih- Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[17] V. J. Leffa, "Clause processing in complex sentences", Proc. 1st International Conference on Language Resources and Evaluation, Granada, Spain, 1998.

김 미 영

e-mail : miykim@sungshin.ac.kr

1995년 3월~1999년 2월 포항공과대학교 컴퓨터공학과 학사

1999년 3월~2005년 8월 포항공과대학교 컴퓨터공학과 박사

2005년 9월~2006년 2월 가톨릭대학교 시간강사

2006년 3월~현재 성신여자대학교 컴퓨터정보학부 전임강사

관심분야 : 자연어처리, 구문분석, 기계번역, 정보검색