문법관계 정보를 이용한 단계적 한국어 구문 분석

이 성 욱†

요 약

본 연구는 한국어 의존 구조를 결정하는 단계적 의존 구조 분석기를 제안한다. 각 단계에서는 주어진 문법관계의 후보열에서 올바른 문법관계를 결정하는데, 대상문법관계의 종류에 따라 독립적으로 수행된다. 문법관계의 후보열은 미리 학습된 지지벡터기계를 이용하여 주어, 목적어, 보어, 부사어 등 7가지의 문법관계로 추정한다. 각 단계에서는 지지벡터기계 분류기와 어절 간의 거리, 교차 구조 금지, 격 제한의 원칙 등의 한국어 언어 특성을 이용하여 대상문법관계를 결정하며, 모든 단계를 거쳐 최종적으로 전체 의존 구조와 문법관계가 결정된다. 트리 및 문법관계 부착 말뭉치를 이용하여 제안된 시스템을 구현 및 실험하였으며 약 85.7%의 정확률을 얻었다.

키워드 : 구문분석, 의존구조, 문법관계, 지지벡터기계

Cascaded Parsing Korean Sentences Using Grammatical Relations

Songwook Lee†

ABSTRACT

This study aims to identify dependency structures in Korean sentences with the cascaded chunking. In the first stage of the cascade, we find chunks of NP and guess grammatical relations (GRs) using Support Vector Machine (SVM) classifiers for all possible modifier-head pairs of chunks in terms of GR categories as subject, object, complement, adverbial, etc. In the next stages, we filter out incorrect modifier-head relations in each cascade for its corresponding GR using the SVM classifiers and the characteristics of the Korean language such as distance between relations, no-crossing and case property. Through an experiment with a parsed and GR tagged corpus for training the proposed parser, we achieved an overall accuracy of 85.7%.

Key Words : Parsing, Dependency Structure, Grammatical Relation, Support Vector Machines

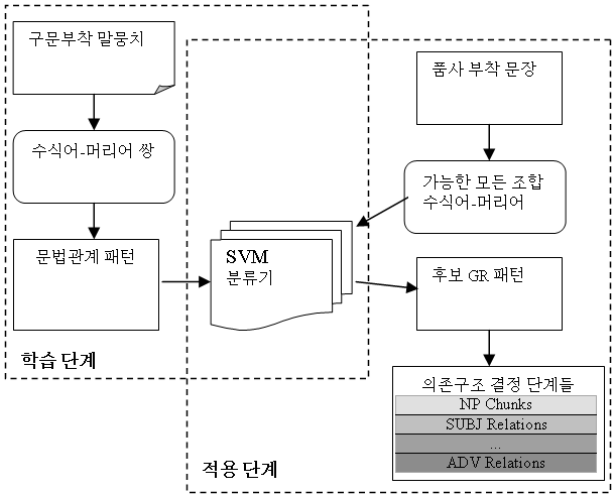
1. 서 론

머리어-수식어 관계를 나타내는 의존 구조 및 문법관계 정보는 정보 검색, 정보 추출, 문서 요약 및 질의응답 시스템 등 대부분의 문서 분석 응용 시스템에 유용하게 이용되어 왔다[1, 2, 3]. 한국어의 의존 구조를 분석하기 위해서 우리는 두 가지 문제를 해결해야 한다. 하나는 '문장에서 어떤 어절이 어떤 수식어의 머리어인가?'하는 문제이며, 다른 하나는 '주어진 수식어와 머리어 사이에 어떤 종류의 문법관계가 성립하는가?'라는 문제이다. 문장의 의존 구조 및 문법관계를 분석하는 연구가 그 동안 많이 수행되어왔는데, [4]에서는 은닉마르코프 모형을 단계적으로 구성하여 문법관계를 결정하였는데, 품사 태깅 방법을 문법관계 결정에 적용하였고, 문법관계를 위한 태거는 어휘 확률과 부모 노드의 분류에 의존하는 문맥 확률을 이용하여 동작한다. [5]는 기억기반 학습법([7])을 이용하여, 단어의 품사정보만을 가지고 명사구 단위화(chunking)와 주어, 목적어를 결정하였고 [6]은 [5]의 방법을 확장하여 문법관계 결정 단계를 단계적으로 덧붙였다. [6]은 먼저 문장을 여러 개의 구로 단위화를 하고 그 다음에 구들 사이의 문법관계를 부착하였다. 그들은 품사 및 어휘 정보를 이용하였고 명사구 및 용언구의 단위화를 수행하여 단위화를 수행하지 않은 방법 보다 더 나은 성능을 얻었다. 그러나 기억기반 학습 알고리즘은 학습 시간보다 실행 시간이 느린 단점과 모든 학습 말뭉치를 저장할 대용량의 기억 공간이 필요한 단점이 있다. [8]은 트리 뱅크 문법 파서에 의해 생성된 문법관계 부착 트리들로 구성된 풍부한 입력을 가정하고, 문법관계를 트리의 구문 요소들에 부착하는 방법을 학습하는데 다시 트리 뱅크를 이용하였다. [9]은 문법관계를 추출하기 위해, 통계적 파서의 출력에 대한 임계값을 조절하였는데 비록 전체 분석 결과의 일부분만을 추출하는 낮은 재현율을 얻었지만, 추출된 문법관계는 높은 정확률을 보였다.

대부분의 이전 연구는 주로 단위화를 먼저 수행하고 각

※ 이 논문은 2007년도 충주대학교 교내학술연구비의 지원을 받아 수행한 연구임.
† 정 회 원 : 국립충주대학교 컴퓨터과학과 전임강사
논문접수 : 2007년 11월 13일, 심사완료 : 2007년 12월 20일

DOI: 10.3745/KIPSTB.2008.15-B.1.69



(그림 2) 다단계 한국어 구문 분석 시스템 구조도

단위(chunk)의 머리어를 탐색하여 구조를 결정하며, 그 이후에 찾아낸 술어-논항 관계에 대한 주어, 목적어, 보어 등의 문법관계를 결정한다. 그런데 [10]은 가능한 모든 문법관계의 쌍을 먼저 찾은 후, 주어진 문법관계 중 올바른 관계를 찾아내는 방법을 주어, 목적어, 부사어 등의 문법관계에 대해 실험하고 제안하였다. [10]은 문장에서 주어진 술어와 명사구에 대한 문법관계 확률을 이용하였고, 어절 간의 거리와 교차구조 금지, 일문일격의 원칙 등의 한국어 특징을 반영하였다. 그러나 [10]의 모형은 술어-논항 관계에만 존재하는 문법관계에 대해서 분석하고 다른 명사구 사이의 문법관계와 술어와 술어 사이의 문법관계에 대해서는 고려하지 않았다.

우리는 단계적 단위화 전략을 이용하여 [10]의 부분 구문분석 방법을 각 단계에 적용하여 의존 구조를 분석할 수 있도록 확장하였으며, 모든 단계를 거친 후에는 모든 문법관계에 대하여 분석할 수 있게 하였다. 각각의 단계에서, 트리부착 말뭉치로부터 자동으로 학습된 지지벡터기계(SVM) 분류기를 각 단계에서 분석하고자하는 문법관계의 결정 모형에 사용하였으며, 자세한 내용은 2장과 3장에서 설명한다.

2. 시스템 구조

(그림 1)은 제안하는 단계적 한국어 구문분석 시스템의 구조도이다. 시스템은 학습 단계와 적용 단계로 나뉜다.

학습단계에서는 모든 의존관계를 구문트리부착 말뭉치에서 추출한다. 추출된 의존관계의 문법관계는 수동으로 부착하였는데 성격에 따라 아래 <표 1>과 같이 7가지로 분류하였다. 추출된 문법관계 쌍들은 문법관계(GR) 패턴들이 되어 SVM분류기의 학습에 사용된다. SVM의 학습은 4장에서 자세히 설명한다.

적용 단계에서는 품사가 부착된 문장이 입력으로 주어지고, 시스템은 가능한 모든 조합의 수식-머리말 관계에 대해 의존관계를 생성한다. 현재 주어진 수식어 m과 머리어 h의

〈표 1〉 문법관계의 종류

문법관계	설명	예제
NP-chunk	Adnominal - NP relation	*한 사람*
SUBJ	Subjective relation	*그가 나오다*
COMP	Complement relation	물이 되다
OBJ	Objective relation	철수를 보았다.
ADV	Adverbial relation	학교에서 나왔다.
VPCONJ	VP - Aux. VP relation VP - VP relation (Complex verb relation) Conjunctive VP-VP relation	*하고 싶다.* *달려 가다.* *더워서 마셨다.*
ADNP	Adnoun Phrase - NP relation	나오는 사람을

Mod Head	-	철수가	학교에서	나오는	것을	보았다
아버지는	-			SUBJ		SUBJ
철수가	-	-		SUBJ		SUBJ
학교에서	-	-	-	ADV		ADV
나오는	-	-	-	-	ADNP	
것을	-	-	-	-	-	OBJ
보았다	-	-	-	-	-	-

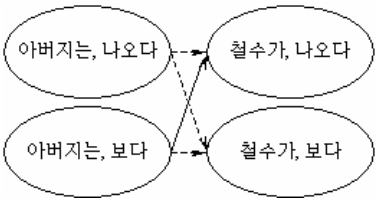
"아버지는 철수가 학교에서 나오는 것을 보았다."
(그림 2) 후보 GR 패턴의 예

문법관계 r에 대한 SVM분류기의 출력을 $SVM_r(m,h)$이라 하자. 모든 가능한 수식어와 머리어 쌍에 대한 문법관계는 $SVM_r(m,h)$이 양의 값일 때, $\arg_r \max SVM_r(m,h)$로 결정할 수 있다. 이 때 결정된 각 쌍의 문법관계 r을 후보 GR패턴이라고 부른다. (그림 2)는 주어진 예문의 후보 GR패턴을 나타낸 것이다.

(그림 2)와 같이 주어진 후보 GR패턴은 의존 구조 결정 단계의 입력으로 주어진다. 의존 구조 결정 단계는 주어진 후보 GR패턴 중에서 올바른 관계와 그른 관계를 구분하는데, [11]의 부분구문 분석 방법을 이용하였다. 다음 3장에서 각 단계에서 올바른 관계를 찾을 수 있는 방법을 설명한다. 의존구조 결정 단계는 문법관계의 각 종류마다 독립적으로 올바른 문법관계를 결정할 수 있도록 7가지 문법관계를 각각 처리하는 7단계로 구성되어있다. 각 문장은 7단계의 의존구조 결정 단계를 거친 후 전체 의존 구조가 결정된다.

3. 각 단계의 의존구조 결정 모형

이 장에서 각 단계에서 문법관계를 분석하는 방법을 설명한다. 우리는 7개의 문법관계를 분석하는 각각의 단계를 두어 최종적으로 각 단계의 결과를 합치는 방법으로 의존 구조를 분석한다. 따라서 각 단계에서는 7가지 문법관계 중 해당하는 하나의 문법관계만을 고려하여 의존 구조를 결정한다. 즉, 주어 관계를 분석하는 단계에서는 목적어나 부사어 등의 다른 관계는 전혀 고려하지 않는다. 현재 단계에서 분석하려는 문법관계를 대상문법관계라고 정의하자. 우리는



(그림 3) 주어에 대한 후보 GR패턴 상태

격 제한 규칙과 교차구조 금지, 의존 거리 등의 언어 특성을 고려하면서, 수식어-머리어 쌍 중에 최대 가중치를 출력하는 쌍들의 열을 찾으면 된다.

다음 알고리즘은 [10]에서 제안한 알고리즘을 모든 의존 구조 관계에 적용할 수 있도록 변형한 것이다. 분석하려는 대상문법관계를 *tgr이라 하고, tgr* 관계에 대한 가중치 함수를 R_{tgr} 이라 하자. 주어진 수식어열 $M_1,...,M_m$ 과 머리어열 $H_1, ..., H_t$ 에 대해 가중치 함수는 수식 (1)와 같다. 수식어와 머리어의 발생은 선후 후보 쌍들에 독립이라고 가정한다. H_k 는 $H_1, ..., H_t$ 중 하나가 i번 째 후보 M_k와 의존 관계를 가지는 것을 의미한다.

$$R_{tgr}(M_1,...,M_m,H_1,...,H_t) \approx \prod_{k=1}^{m} SVM_{tgr}(M_k,H_k) \cdot f(M_k,H_k,M_{k-1},H_{k-1}) \cdot P(d \mid r=tgr) \quad (1)$$

$$f(M_k,H_k,M_{k-1},H_{k-1}) = \begin{cases} 0, & \text{iff } (loc(\mathrm{H}_k) = loc(H_{k-1}) \text{ and } H \in \text{verb}), \quad (2) \\ 1, & or\ (loc(M_k) < loc(H_{k-1}) < loc(H_k)) \quad (3) \\ & \text{otherwise} \end{cases}$$

대상문법관계가 주어일 때(tgr=subj), (그림 2)에 표시된 주어 후보 GR 패턴의 가능한 모든 상태들은 (그림 3)과 같이 나타낼 수 있다. 다단계 모형의 각 단계의 목적은 식(1)을 최대화하는 최적 경로를 대상문법관계의 후보 GR패턴에서 찾는 것이다.

식 (2)에서 $loc(\mathrm{H}_k) = loc(H_{k-1})$ 은 (그림 3)의 평행한 두 개의 점선처럼, 어떤 용언이 선행하는 두 개 이상의 명사구와 동일한 문법관계를 중복해서 가지는 것을 뜻하며, 이 때, 격 제한 원칙을 반영하여 가중치를 0으로 만든다. 식(3)의 $loc(M_k) < loc(H_{k-1}) < loc(H_k)$ 은 (그림 3)의 대각으로 내려오는 점선처럼 이웃하는 구조와 교차구조가 만들어진 것을 나타내며, 식(3)은 이러한 경로의 가중치를 0으로 하여 제거한다. 비터비 알고리즘[11]을 이용하면 식(1)을 최대화하는 최적경로를 쉽게 찾을 수 있다.

명사구와 용언구 사이의 거리 정보도 둘 사이에 관계를 가지는 데 큰 영향을 끼친다. 거리 정보는 두 단어의 관계에 있어서 중요한 정보임이 밝혀져 왔다. 식(1)에서는 어절 간 거리 정보 d를 특정 문법관계의 거리 분포 P(d|r=tgr)로 반영하였다. 식(1)의 $SVM_{tgr}(M_k,H_k)$ 은 지지벡터기계를 이용하여 추정하는데 이는 다음 장에서 설명한다.

4. 지지벡터 기계를 이용한 문법관계 학습

주어진 수식어와 머리어에 존재하는 문법관계에 대한 확률은 MLE (maximum likelihood estimation) 등의 방법을 사용하면 추정할 수 있으나 어휘 정보의 사용에 따른 자료부족 문제가 발생한다. 그래서 우리는 SVM[12]를 이용하여 주어진 수식어와 머리어에 존재하는 문법관계에 대한 가중치를 계산한다. 명사구, 명사구의 조사, 용언 등의 어휘 자질과 각 어휘의 품사 자질을 사용하여 SVM을 학습한다. 자질벡터의 차원은 각 자질의 어휘의 개수의 총합이 되며 각각의 자질은 자질의 유무에 따라 이진값으로 표현되었다. 학습데이터에서 출현하는 각 수식어와 머리어사이의 문법관계가 그 문법관계를 위한 SVM의 양의 자질로 사용되었고 동시에 다른 문법관계를 위한 분류기의 학습에는 음의 자질로 사용되었다.

반복 실험 결과, SVM의 커널은 시스템 성능에 큰 영향을 끼치지 않아 선형 커널을 사용한다. SVM은 이진 분류기이므로 각 문법관계에 대한 분류기를 각각 학습하였고 실험에는 SVMlight[15]를 이용하였다.

5. 실험결과

우리는 구문구조가 부착된 한국어정보베이스 말뭉치[13]을 실험에 사용했다. 실험에는 145,630어절의 11,932문장을 사용했다. 이 말뭉치로부터 용언 및 수식 명사구 쌍 120,830개에 수동으로 문법관계를 부착하고 SVM의 문법관계 학습에 사용하였다. 학습에서 사용되지 않은 5,056어절의 475문장을 평가집합A로 사용하였고, 다른 구문 분석기와의 비교를 위해서 195개의 문장으로 이뤄진 평가집합B를 사용하였다. 정확률과 재현율로 평가하고 F1 평가-2*P*R/(P+R) 로도 나타냈다. <표 2>는 제안 시스템의 평가 집합에 대한 성능을 나타낸다.

비단계적 방법의 결과는 단계적 방법을 적용하지 않고 각 문법관계를 따로 분석하지 않고 한번에 모든 관계를 분석했을 때의 결과이다. <표 2>와 같이 단계적 의존구조 분석 방법이 더 나은 결과를 보였다. <표 2>는 전체 구문 분석에 사용된 이공주[14]의 방법과 단계적 구문분석 방법을 이용한 본 시스템의 성능을 비교하고 있다. [14]의 방법은 구-구조 트리를 출력하므로 그 결과를 의존 구조로 변형하였으며 평가집합B를 이용하였다. <표 2>와 같이 제안 시스템의 성능이 [14]의 방법보다 약간 높은 결과를 보였으며, 이러한 결과는 제안하는 부분 구문 분석 방법을 단계적으로 적용한

〈표 2〉 제안 시스템의 성능

평가 집합	방법	Avg. P	Avg. R	Avg. F1
A	비단계적 방법	82.0	81.5	81.8
	단계적 방법	85.9	85.3	85.7
B	제안 시스템	86.3	85.4	85.9
	이공주[14]모형	86.1	82.7	84.4

방법이 일반적인 통계적 구문 분석 기법의 성능에 필적한다고 할 수 있다. 결과적으로 의존구조의 중의성을 단계적 문법관계 분석 방법을 통해 효과적으로 해소되었다고 할 수 있다.

대부분의 오류는 GR 후보 패턴을 결정하는 과정에서 발생하는 오류이며 이것은 시스템의 언어 특성 반영과 상관없이 문법관계의 수식어와 머리어의 문법관계 결정에서 발생하는 오류이다. 대부분 이런 오류는 자료 부족 문제에 기인한다.

6. 결론 및 향후 과제

본 연구에서 한국어 문법관계를 결정하는 단계적 의존 구조 분석기를 제안하고 구문구조 말뭉치를 이용하여 구현하였다. 제안된 구문 분석기는 7가지 문법관계에 대해 각각 독립적으로 분석하여 최종적으로 하나의 의존 구조를 결정하며 각 의존 구조의 문법관계도 동시에 결정된다. 문법관계에 대한 통계 정보는 구문구조와 문법관계 부착 말뭉치에서 추출하였고 지지벡터 분류기를 학습하는데 사용하였다. 제안된 방법은 수식어와 머리어 사이에 존재하는 문법관계에 대한 통계적 정보를 각 문법관계의 중의성 해소에 단계적으로 사용하여 전체 문장의 의존 구조를 결정하였으며, 한국어의 언어특성인 교차구조 제한, 격제한 원칙 및 어절간의 거리 등을 각 문법관계의 중의성 해소에 이용하였다. 실험을 통해 의존 구조 분석에서 약 85.7%의 정확률을 얻었다. 더 신뢰할 만한 결과와 더 나은 성능을 위해서 좀더 많은 데이터가 필요하며, 현재 제안된 방법과 대용량의 말뭉치에서 추출한 공기정보와 기계가독사전을 이용한 단어의미분별 모형과 결합하는 방법을 연구하고 있다.

참 고 문 헌

[1] Grenfenstette, G. “SQLET: Short query linguistic expansion techniques, palliating one-word queries by providing intermediate structure to text”, In Proc. of the RIAO'97, pp.500-509, 1997.

[2] Palmer, M., Passonneau, R., Weir, C. & Finin, T. “The KERNEL text understanding system”, Artificial Intelligence, Vol. 63, pp.17-68, 1993.

[3] Yeh, A. “Using existing systems to supplement small amounts of annotated GRs training data”, Proc. of the ACL2000, pp.126-132. Hong Kong, 2000.

[4] Brants, T., Skut, W. & Krenn, B. “Tagging grammatical functions”, In Proceedings of the 2nd Conference on EMNLP, pp.64-74. Providence, RI., 1997.

[5] Argamon, S., Dagan, I. & Krymolowski, Y. “A memory-based approach to learning shallow natural language patterns”, In Proceedings of the 36th Annual Meeting of the ACL, pp.67-73. Montreal, Canada, 1998.

[6] Buchholz, S., Veenstra, J. & Daelemans, W. “Cascaded GR assignment”, In Proceedings of the Joint Conference on EMNLP and Very Large Corpora, pp.239-246, 1999.

[7] Stanfill, C. & Waltz, D. “Toward memory-based reasoning”, Communications of the ACM, Vol. 29, pp.1213-1228, 1986.

[8] Blaheta, D. & Charniak, E. “Assigning function tags to parsed text”, In Proceedings of the 1st Conference of the NAACL, pp.234-240. Seattle, WA, 2000.

[9] Carroll, J. & E. Briscoe “High precision extraction of GRs”, In Proceedings of the 19th International Conference on Computational Linguistics (COLING), pp.134-240, Taipei, Taiwan, 2002.

[10] 이성욱, 서정연, “한국어 문법관계에 대한 부분구문분석”, 정보과학회논문지 소프트웨어 및 응용, 제32권10호 지 pp.984-989, Oct. 2005.

[11] Viterbi, A. J. “Error bounds for convolution codes and an asymptotically optimal decoding algorithm”, IEEE trans. on Information Theory, Vol. 12, pp.260-269, 1967.

[12] Vapnik, V. N. “The Nature of Statistical Learning Theory”, Springer, New York, 1995.

[13] Lee, K.-J., KIM, J.-H., Choi, K.-S. & Kim, G. C. “Korean syntactic tagset for building a tree annotated corpus”, Korean Journal of Cognitive Science, Vol. 7, No. 4, pp.7-24, 1996.

[14] Lee, K.-J., Kim, J.-H., & Kim, G.-C. “An Efficient Parsing of Korean Sentence Using Restricted Phrase Structure Grammar”, Computer Processing of Oriental Languages, Vol.12, No. 1, pp. 49-62, 1997.

[15] http://svmlight.joachims.org

이 성 욱

e-mail : leesw@cjnu.ac.kr

1996년 서강대학교 전자계산학과 학사

1998년 서강대학교 컴퓨터학과 석사

2003년 서강대학교 컴퓨터학과 박사

2003년~2004년 서강대학교 산업기술연구소 연구원

2003년~2005년 서강대학교 정보통신대학원 대우교수

2004년~2005년 LG전자 기술원 선임연구원

2005년~2007년 동서대학교 컴퓨터공학과 전임강사

2007년~현재 국립충주대학교 컴퓨터과학과 전임강사

관심분야 : 형태소 및 구문 분석, 단어의미분별, 대화 언어처리, 한국어 생성 등.