구문 그래프를 이용한 구문적 애매성 분석

김 재 훈, 서 정 연, 김 길 창
한국과학기술원 전산학과

Analysis of Structural Ambiguities Using Syntactic Graph

Jae-Hoon Kim, Jungyun Seo, Gil Chang Kim
Dept. of Computer Science, KAIST

요 약

한국어는 그 자체의 특성 때문에 영어와는 또 다른 형태의 구문적인 애매성을 포함하고 있다. 이와 같은 구문의 애매성을 해결하기 위해서는 여러 가지의 정보가 필요할 것이다. 예를 들면, 품사정보의 세분류, 명사들의 의미 속성정보들이 그것이다. 본 논문은 한국어 문장의 구문적인 애매성을 해결하기에 앞서 먼저 한국어 문장에 어떤 형태의 애매성이 포함되어 있는 지를 조사.분석한 것이다. 본 논문에서는 구문적인 애매성을 효율적으로 분석하기 위한 수단으로 구문 그래프를 이용하였다.

한국어 문장에는 다품사에 의한 애매성, 조사구 부착에 관한 애매성, 복합 체언구에 관한 애매성, 부사구 부착에 관한 애매성, 관형어의 수식 범위에 관한 애매성이 있다. 이들 중에서 복합 체언구에 의한 애매성이 가장 많은 애매성을 가지고 있었다. 즉, 실험 대상 문장에서 발생가능한 전체의 애매성의 62%가 복합체언구에 관한 것이다. 따라서 한국어에서는 복합체언구에 관한 구문 구조적인 애매성 해소가 가장 우선적으로 해결해야 할 과제이다.

I. 서론

주어진 하나의 대상에 대해서 하나 이상의 해석이 가능할 때, 그 대상에 대해서 애매성이 있다고 말한다. 자연어 처리에서는 주어진 문장에 대해서 여러 형태의 애매성을 제거하여 하나의 해석을 추출하는 것이 가장 기본적인 문제이다. 이와 같은 문제를 애매성 해소 혹은 해석이라고 한다. 애매성은 자연언어 처리의 모든 단계에서 발생하기 때문에 각 단계에서 처리가능한 모든 애매성을 해소해야 한다. 특히 본 논문에서는 여러 가지의 애매성 중에 구문적인 애매성에 관해서 다루고자 한다. 구문적인 애매성으로 영어에서 가장 대표적인 것은 PP-attachment(

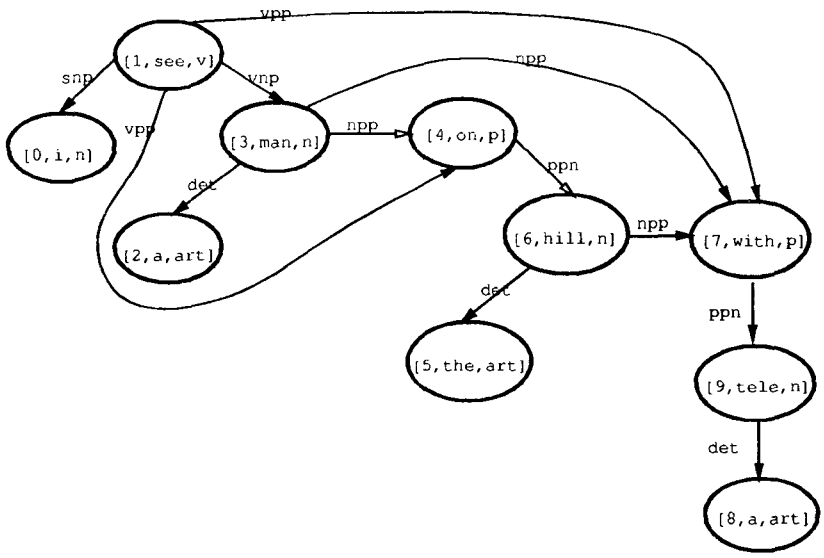

그림 1: *I saw a man on the hill with a telescope*의 구문 그래프

전치사 부착)이며, 이를 해소하기 위한 많은 연구가 진행되어 왔고, 또 아직도 진행되고 있다[1, 2, 3, 4, 5]. 그러나, 한국어의 경우에서 애매성 해소에 관한 연구는 물론이거니와 어떤 종류의 애매성들이 존재하는 지에 관한 연구도 거의 없다. 본 논문은 구문적인 애매성을 해소하기 위한 전단계로써 우선 한국어에서는 어떠한 형태의 구문적인 애매성이 있는 지를 분석하고자 하는데 그 목적이 있다.

작은 조각들이 모여서 하나의 큰 덩어리를 이룬다는 것이 문장 구성의 가장 기본적인 원리이다. 문장을 구성하는 작은 조각들이 큰 덩어리를 어떻게 구성하는 지에 관한 표현 방법으로는 head-modifier 방법, immediate constituent 방법, slot-filler 방법이 있다[6]. 본 논문에서는 head-modifier 방법의 일종인 구문 그래프(syntactic graph)를 이용하여문장의 애매성을 분석하고자 한다. 구문 그래프[7, 8]는 triple의 집합으로 구성되며, 각 triple은 label, 중심어, 그리고 수식어로 구성되었다. Label은 해당 triple을 생성할 때 사용된 문법 규칙에 따라 유일하게 정해진다. 예를 들면, $S \rightarrow NP + VP$라는 문법 규칙에 의해서 생성된 triple은 *snp*라는 label명을 갖는다. 그림 I.은 구문 그래프의 예이다.

구문 그래프는 문장에 사용된 한 단어는 하나의 해석에서는 둘 이상의 다른 단어에 의해서 수식되어질 수 없다는 성질을 이용해서 그래프 상에 구문적인 애매성을 잘 표현할 수 있다. 즉, 그래프 상에서 여러 개의 in-arc를 갖는 node는 여러 단어로부터 수식을 받고 있음을 나타내며, 이 단어를 ambiguous point라고 한다. 구문 그래프로부터 가능한 구문해석을 구할 수 있다. 그림 I.에서는 *on*의 in-arc가 2개, *with*의 in-arc가 3개이다. 따라서 전체 가능한 구문 해

석의 수는 2 × 3 = 6개가 될 것이다. 그러나, 위의 실질적으로는 5가지의 해석이 가능하다. 이것은 구문 그래프만으로는 어떤 구문해석이 불가능한 지를 알 수 없다. 이와 같이 실질적으로 일어날 수 없는 triple들에 대한 정보를 저장해야 하는데, 이것을 exclusion matrix라고 한다.

구문 그래프의 ambiguous point를 조사.분석함으로써 문장의 애매성을 쉽게 분석할 수 있기 때문에 본 논문에서 구문 그래프를 이용하였다.

제2장에서 한국어 문장의 애매성을 분석하기 위한 기본적인 가정과 환경에 대해서 설명하고, 제3장에서 한국어 문장의 애매성을 분석하고 제4장에서 결론을 맺고자 한다.

II. 연구 환경 및 가정

본 논문는 가능한 모든 한국어 문장의 구조적인 애매성을 분석할 수 없기 때문에 몇 가지의 가정을 기반으로 이루어 졌다.

첫째로 한 문장 내에서 발생하는 구문적인 애매성만 고려했다.

둘째로 구문요소화[11]를 가정하였다. 즉, 여러 개의 형태소가 하나의 문법 범주(syntactic category)에 속할 때, 그들을 하나의 덩어리로 묶어서 하나의 문법 범주로 간주하는 것을 말한다. 대표적인 구문요소화는 용언이다. 용언의 구문요소화는 보조 용언, 보조적 연결어미를 하나의 동사로 간주한다. 예를 들면, "두 사람이 묵을 방 하나를 예약하고 싶습니다"에서 "예약하고 싶"을 하나의 동사로 간주한다는 것이다. 이렇게 하는 이유는 형태소 해석에서는 이들을 분리하나, 사실상 이들은 본 용언의 의미를 보조하며 위치적으로도 본 용언 바로 다음에 온다. 따라서 이들을 하나의 문법 범주로 간주한다.

세째로 단어의 품사 정보만을 이용하여 구문적인 애매성을 분석하였다. 언어학에서는 한국어의 단어 품사를 일반적으로 8품사로 분류한다[9]. 그러나, 본 논문에서는 조금 더 세분하였다[10]. 예를 들면, 명사는 보통명사, 의존명사, 고유명사로 나뉘었다. 또, 어미도 하나의 문법 범주로 간주하였다. 어미는 어간과 결합하여 문장 내의 다른 말과의 문법적 관계를 결정한다. 이와 같은 의미에서 어미도 조사와 마찬가지로 하나의 문법 범주로 간주하는 것이 표현의 통일성을 잃지 않을 것이다. 특히, 문장의 구조를 중심어와 수식어 관계가 고려될 경우, 병렬문에 대해서는 대등적 연결어미를 하나의 문법범주로 간주해야만 한다. 예를 들어, 예문 1의 분석은 그림 2과 같이 이루어져야만 한다.

여름은 덥고 겨울은 춥다. (1)

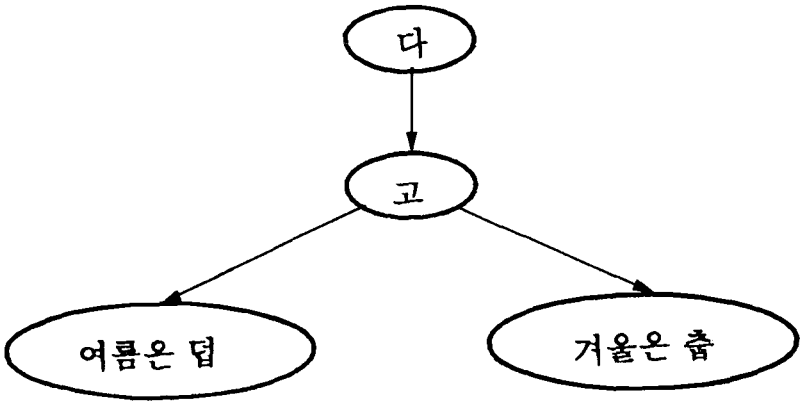

그림 2: 대등적 연결어미에 대한 문장 분석

네째로 단어 자체의 의미적인 애매성은 고려 대상에서 제외한다. 즉, 단어 자체의 품사의 다중성은 허락하나, 같은 품사에 대해서 다의성을 허락하지 않는다.

위에서 설명한 가정을 기준으로 하여 박인씨가 지은 "호텔 영어회화"citeref:Park의 unit 1, 호텔 예약에 관한 대화를 중심으로 44개의 문장을 추출하여 분석하였다.

III. 구문적 애매성

1 다품사

영어에서와 마찬가지로 다품사에 관한 애매성은 한국어에서도 포함되어 있다. 예를 들어, 단어 "몇"은 수사이면서 관형사이다. 따라서, (2)는 그림 3와 같이 해석되어 "몇"를 유발시킨 지를 구문 그래프를 이용하여 쉽게 알 수 있다. 즉, "몇"이 ambiguous point가 된다.

자녀 몇 분이 함께 오십니까? (2)

이하의 설명에서는 구문 그래프를 지면 관계상 괄호 표기에 의해서 표현한다.

2 조사구 부착(PP-attachment)

PP(particle phrase)-attachment는 한국어에서 주로 용언의 범위(predicate coverage)에 의해서 주로 발생된다. 여기서 PP는 격조사(관형격 조사는 생략)나 보조사가 체언이나 용언의 명사형 아래에서 그들을 이끌 때를 말한다. 따라서 PP는 문장 전체를 꾸미는 경우이거나 용언

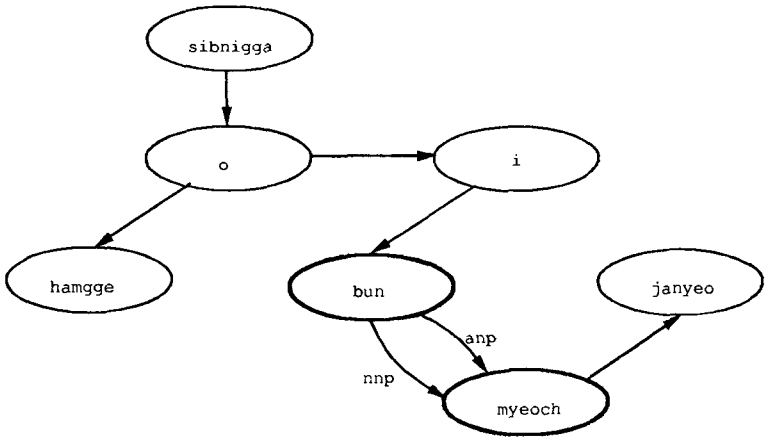

그림 3: 다품사에 의한 애매성

을 꾸미는 경우이다. 그러나, 많은 경우에 용언을 꾸미기 때문에 여기서는 이것에 대해서만 살펴 보기로 한다. (3)에서 PP "두 개의 침실이"는 구문 구조상으로 용언 "있다"를 꾸밀 수도 있고(4), 또 다른 용언 "부탁하다"를 꾸밀 수도 있다(5). 여기서의 올바른 해석은 (4)이다.

그들이 나에게 두 개의 침실이 있는 스위트로 예약을 부탁했읍니다. (3)

그들이 나에게 ((두 개의 침실이) 있)는 스위트로 예약을 부탁했읍니다. (4)

그들이 나에게 ((두 개의 침실이) (있는 스위트로) (예약을) 부탁했)읍니다. (5)

3 복합 체언구

복합 체언구는 상당히 복잡한 구조를 가진다. 먼저, 체언구가 체언구를 수식할 경우이다(N-N-N ...). (6)에서 "다음"의 중심어는 "주"가 되어야 할지 "수요일"이 되어야 할지를 구문 구조상으로는 결정할 수 없다. 따라서, (6)는 (7)와 (8)와 같은 구조로 해석된다.

다음 주 수요일에 더블룸으로 예약하고 싶은데요. (6)

((다음 주) 수요일)에 더블룸으로 예약하고 싶은데요. (7)

((다음) (주 수요일)에 더블룸으로 예약하고 싶은데요. (8)

(N-N-N ...)과 같은 복합체언구는 항상 용언에 의해서 또 다른 형태의 애매성을 가지고 있다(N-N ... VP). 왜냐 하면, 일반적으로 한국어 문장에서는 특히 대화체 문장에서는 체언, 그 자체만으로도 용언을 수식할 수 있기 때문이다. 이 경우에는 일반적으로 조사가 생략된 경우가 대부분이다. (9)는 (6)가 (N-N ... VP) 구조로 해석될 경우이다.

((다음) (주 수요일에) (더블룸으로) 예약하고_싶)은데요. (9)

두번째로 접속사(접속 부사(혹은, 또는, 및, ...), 접속 조사(과/와, 하고, ...))에 의해서 연결된 접속 복합 체언구이다(NP-conj-NP-conj-NP ...). 접속조사가 체언구를 서로 연결하여 문장에서 하나의 복합체언구를 형성하게된다. 이 때에 접속조사들사이의 중심어 결정에 있어서 발생된 애매성이다. 사실상 이 경우에는 뒤에 나오는 접속조사를 중심어로 할 경우에는 이와 같은 형태의 애매성은 발생하지 않을 것이다. (10)의 복합 체언구 "전화 번호와 이름과 주소"에서 체언 "이름"은 "와"에 수식되는 것으로 해석할 수도 있고(11), 또 "와"에 수식되는 서식되는 것으로 해석할 수도 있다(12). 이 경우는 (N-N-N ...) 구조와는 달리 용언을 수식하는 해석은 일어나지 않는다.

전화 번호와 이름과 주소를 말씀해 주시겠습니까? (10)

((전화 번호)와 ((이름)과 (주소)))를 말씀해 주시겠습니까? (11)

((전화 번호)와 (이름))과 (주소))를 말씀해 주시겠습니까? (12)

세번째로는 접속 복합 체언구가 문두에서 나오면서 접속사가 생략될 경우이다(NP-,-NP-conj-NP ...). 한국어 문장은 접속사가 생략된 자리에 반드시 쉼표(comma)를 사용해야 한다[9, page 170]. 일반적으로 문두에서 체언구 다음에 쉼표가 나오는 경우는 독립어와 혼란이 일어난다. (13)는 (10)의 예문에서 접속 조사 "와"가 생략된 형태이다. 이 때에 (14)의 "전화 번호"는 독립어로서 문장 전체를 수식하고, (15)의 "전화 번호"는 앞에서 설명한 접속 복합 체언구의 구성 요소로 사용되어 접속사 역할을 하는 쉼표(,)를 형태적으로 꾸미고 있다.

전화 번호, 이름과 주소를 말씀해 주시겠습니까? (13)

((전화 번호,) (이름과 주소를 말씀해 주시겠) 습니까)? (14)

((전화 번호, 이름과 주소)를) 말씀해 주시겠 습니까? (15)

4 부사구 부착(Adverbial phrase attachment)

부사는 주로 용언 앞에서 그 용언을 꾸밈으로써 그 의미를 더욱 분명히 해 주는 말을 말한다[9, page 137]. 그러나, 영어에서와는 달리 체언, 용언, 수식언(관형사, 부사) 앞에서 그들의 뜻을 한정하는 역할도 가지고 있다. 부사의 이와 같은 성질 때문에 부사의 부착 문제가 발생하게 된다. (16)의 구문 구조적인 해석은 (17), (18), (17)이 가능하다. (17)은 부사 "매우"가 용언(형용사) "조용하다"를 수식하는 경우이고, (18)는 부사 "매우"가 체언(명사) "방"를 수식하는 경우이며, (19)는 부사 "매우"가 용언(동사) "주다"를 수식하는 경우이다.

매우 조용한 방으로 주십시요. (16)

((매우) 조용하)ㄴ 방으로 주십시요. (17)

((매우) ((조용한) 방)으로 주십시요. (18)

((매우) (((조용한) 방)으로) 주시)ㅂ시요. (19)

5 관형어의 수식 범위

관형어는 다음에 오는 체언을 꾸미는 말로 관형사, 체언+관형격 조사, 체언+서술격 조사의 관사형, 용언의 관형사형, 용언의 명사형+관형격 조사의 형태를 갖는다. 관형어 다음에 나오는 복합 체언구일 때에 관형어는 ambiguous point가 된다. (20)의 관형어 "그"는 체언 "날짜"를 수식하는 경우(21)와 체언 "숙박"을 수식하는 경우(22)와 체언 "시설"을 수식하는 경우가 있다.

그 날짜에 그만한 숙박 시설이 가능한지 컴퓨터로 확인하는 ... (20)

((그) 날짜)에 그만한 숙박 시설이 가능한지 컴퓨터로 확인하는 ... (21)

((그) (날짜에 그만한) 숙박) 시설이 가능한지 컴퓨터로 확인하는 ... (22)

((그) (날짜에 그만한 숙박) 시설)이 가능한지 컴퓨터로 확인하는 ... (23)

I	II	III_a	III_b	III_c	III_d	IV	V
4	27	87				10	13
		21	61	2	3		

표 1: 애매성의 빈도수

IV. 평가 및 결론

한국어 문장에서 가장 많이 발생되는 애매성은 복합 체언구에 의 해서 발생되는 것이다(약 62%). 물론 44개의 실험 대상 문장만으로 한국어 전체에서 발생되는 애매성의 62%가 복합 체언구에 의해서 발생될 것이라고 말할 수 없다. 이 자료는 단지 본 논문에서 실험 대상으로 추출된 것에 대해서 얻어진 자료이다. 표 1은 실험에서 분석된 애매성의 빈도수를 표식화 한 것이다.

표 1에서 I[1], II[2], III_a[3], III_b[4], III_c[5], III_d[6], IV[7], V[8]은 애매성의 종류를 나타내고, 표내의 숫자는 각 애매성의 발생빈도를 나타낸다.

44개의 문장 만으로 분석되었기 때문에 한국어 문장의 모든 애매성이 분석된 것은 아니므로 더 많은 문장을 분석하면 또 다른 형태의 애매성이 존재할 것 같다.

앞으로는 본 논문을 토대로 해서 한국어의 구문 구조적인 애매성의 해결 방법을 하나 하나 밝혀야 할 것이다.

참고 문헌

[1] S. I. Small, G. W. Cottrell and M. K. Tanenhaus, *Lexical Ambiguity Resolution: Perspective from Psycholinguistics, Neuropsychology, and Artificial Intelligence*, Morgan Kaufmann Publishers, Inc., 1988

[1]다의성
[2]조사구 부착
[3]복합체언구(N-N-N ...)
[4]복합체언구(N-N ... VP)
[5]복합체언구(NP-conj-NP-conj-NP ...)
[6]복합체언구(NP-,-NP-conj-NP ...)
[7]부사구 부착
[8]관형어의 수식 범위

[2] K. Nagao, Dependency Analyzer: A Knowledge-Based Approach to Structural Disambiguration, *Coling-90*, pp. 282-287, 1990

[3] K. Jensen and J.-L. Binot, Disambiguation Prepositional Phrase Attachments by Using On-line Dictionary Difinitions, *Amer. J. of Computational Linguistics*, vol. 13, no. 3-4, pp. 251-260, 1987.

[4] K. Nagao, Constraints and Preferences: Integrating Grammatical and Semantic Knowledge for Structural Disabiguation *Proc. of PRICAI*, pp. 484-489, 1990.

[5] Y. Wilks, X. Huang, and D. Fassm, Syntax, Preference and Right Attachment *Proceedings of Int'l Joint Conf. on Artificial Intelligence*, pp. 779-784, 1985.

[6] T. Winograd, *Language an s Cognitive Process*, Addison-Wesley publishing Company, 1983.

[7] J. Seo and R. F. Simmons, Syntactic Graphs: A Representation for The Union of All Ambiguous Parse Trees, *Amer. J. of Computational Linguistics*, vol. 15, no. 1, pp. 19-32, 1989.

[8] H.-C. Rim, J. Seo, and R. F. Simmons, Transforming Syntactic Graphs into Semantic Graphs, *Proc. of the 28th Annual Meeting of the Assoc. for Computational Linguistics*, pp. 47-53, June 1990.

[9] 조규빈, 하이라이트 고교 문법, 지학사, 1986.

[10] 김재훈, 서정연, 한국어의 구문적 애매성 분석 - Syntactic graph를 이용하여 -, 한국과학기술원, 전산학과, 컴퓨터 시스템 실험실, 내부 메모, 1992.

[11] 안동언, 기계번역을 위한 한국어 해석에서 형태소로부터 구문요소의 형성에 관한 연구, 한국과학기술원, 전산학과, 석사학위논문, 1987.

[12] 박인, 호텔 영어 회화, 도서출판 소명, 1988.