언어지식 획득 과정에서의 수렴성 보장에 관한 연구

이현아*, 박재득, 박동인
시스템공학연구소 자연어정보처리연구부

{halee, jdpark, dipark}@seri.re.kr

Researches on the Convergence of Linguistic Knowledge Acquisition Process

Hyun-A Lee, Jay-Duke Park, Dong-In Park
Natural Language Information Processing Department, Systems Engineering Research Institute

요 약

다양한 응용 목적의 대규모 실용적 언어지식 구축을 위해서는 한국어의 모든 언어현상을 수용할 수 있는 이상적인 언어지식(optimal linguistic knowledge) 획득을 목표로 연구해 나가야 한다. 본 연구에서 언어지식의 획득은 주어진 말뭉치의 분석을 통해 이루어진다. 주어진 말뭉치에서 새로운 언어현상이 발견되었을 경우, 기존의 언어지식은 새로운 언어현상을 수용할 뿐만 아니라 기존에 발견되었던 언어현상도 함께 수용할 수 있도록 바뀌어져야 한다. 이러한 변화의 원칙이 보장되어야만 언어지식의 양적 확장과 함께 질적 확장을 이룰 수 있다.

본 연구에서는 언어지식의 질적 확장을 언어지식의 수렴성이라고 정의하고 수렴성 보장을 위한 방법론을 연구한다. 수렴성 보장을 위해서는 먼저 언어지식 획득과정이 공정화, 자동화되어야 하고 언어지식이 변화할 때 수렴을 확인하는 과정이 필요하다. 수렴을 확인하기 위하여 구문구조 데이터베이스와 역사전(Inverted Dictionary)을 이용하는 방법을 제안한다. 지금까지는 언어지식의 양적 확장에만 치중해 왔으나 본 연구에서 제안된 방법으로 언어지식이 구축된다면 질적 확장도 함께 도모할 수 있을 것으로 기대된다.

1 서론

인간이 가지거나 표현하는 대부분의 지식이 언어의 형태로 기록되므로 자연어를 이해하는 컴퓨터만이 모든 정보에 접근할 수 있고, 복잡한 시스템에 모든 사람이 손쉽게 접근할 수 있기 위해서는 자연어 인터페이스가 필요하다[Allen95]. 이처럼 다양한 분야에서 자연어 처리가 요구됨으로써 자연어 처리에 필수적인 언어지식 구축에 점점 더 많은 관심이 고조되고 있다.

지금까지는 언어지식과 관련해서 저장구조나 정보 접근 방식 등을 중점적으로 연구해 왔고[이승 94] 구축 자체에 관해서는 잘 정의된 방법 없이 양적 확장에만 치중해 온 것이 사실이다. 양적 확장에만 치중해 온 결과 언어지식의 질적 확장은 제대로 되지 않아 여러 연구자들이 믿고 공유할 수 있는 대규모 고품질의 언어지식이 아쉬운 상황이다.

언어지식의 질적 확장을 도모하기 위해서는 언어지식 구축과정에 잘 정의된 방법이 적용되어야 하고[이현 96] 무엇보다도 양적으로 늘어감에 따라 질적으로도 수렴할 수 있게 하는 방법의 연구가 필요하다. 따라서 본 논문에서는 언어지식의 질적 수렴을 보장하는 언어지식 구축 방법에 관하여 연구하였다.

2 언어지식의 점진적 수렴성 정의

언어지식의 수렴은 어떤 주어진 말뭉치를 분석하는 과정에서 새로운 언어현상(문법현상)을 발견해 나가는 과정과 같다고 할 수 있다. 말뭉치에 나타나는 언어현상이 새로운 것이 아니라면 기존의 언어지식이 변경될 필요가 없고 새로운 언어현상이 나타날 때 언어지식은 변경되어야 하며 변경된 언어지식은 기존의 언어지식보다 더 높은 언어현상의 수용력을 가지면서 이상적인 언어지식에 좀더 가까이 수렴하기 때문이다.

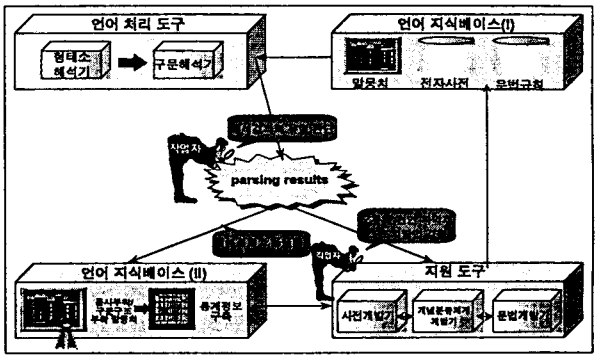
그러나 변경된 언어지식이 항상 더 높은 언어현상의 수용력을 가진다고 보장할 수 없다. 새로운 언어현상의 수용을 위해 변경된 부분이 기존에는 수용하던 언어현상을 수용할 수 없게 만드는 경우가 있을 수 있기 때문이다. 이것은 언어지식 획득에 있어서 잘 정의된 방법론의 부재로 인하여 현단계에서의 언어지식의 변화가 이전 단계에 영향을 미치는지에 대한 여부를 확인하지 않고 무작정 변화시키기 때문이다. 이렇게 수렴에 대한 여부가 보장되지 못하는 경우 언어지식이 양적 확장은 이루어지지만 질적 확장은 이루어지기가 어려울 것이다. 따라서 본 연구에서는 언어지식을 획득하는데 있어서 질적 확장을 보장하는 방법을 연구하였다.

3 수렴성 보장에 관한 연구

본 연구에서 다루는 언어지식은 구문 단계까지의 지식으로 이후 구문지식이라고 칭한다. 구문지식 획득 과정에서의 구문지식의 점진적 수렴성을 보장하기 위해서는 첫째, 새로운 언어현상에 대하여 신속하고 정확하게 구문지식을 수정할 수 있도록 해야 하고 둘째, 현단계에서의 수정이 이전 단계에 영향을 미치는지에 대한 여부를 확인할 수 있어야 한다. 이를 위하여 구문지식 획득 과정의 공정화와 수렴성 확인을 위한 방법론이 필요하다.

3.1 구문지식 수정

컴퓨터에 의해 수행되는 작업은 항상 일관성이 있고 공정화되어 있는 반면 사람이 개입되는 작업은 사람마다 또는 시간이 지남에 따라 일관성이 결여될 수 있으므로 본 연구에서는 이러한 작업들을 공정화하고자 한다. 전단계 연구인 [Lee97]에서 제안된 구문지식 획득 과정은 [그림 1]과 같다. [Lee97]에서는 구문지식 획득 과정에서 사용되는 각종 지원도구와 통합 환경을 마련하여 전체적인 과정에 어느 정도 공정화와 자동화를 이루었다.



[그림 1] 구문지식 획득의 흐름도[Lee97]

[그림 1]을 보면 구문 해석 결과를 보고 판단하여 그 결과에 따라 적절한 조치를 취하는 것까지가 바로 사람이 개입되는 작업이다. 이 과정을 좀더 공정화하기 위하여 구문 해석 결과의 유형을 정의하고 어떤 유형에 속하는지를 판단하여 유형에 따른 조치 방법을 결정하는 기준을 [표 1], [표 2]와 같이 제시한다.

[표 1] 문장 해석 결과의 유형[시스 97]

해석결과 \ 문장유형		정문	비문
해석성공	맞음	PASS	NULL
	틀림	*CASE1*	*CASE2*
해석실패	맞음	NULL	PASS
	틀림	*CASE3*	*CASE4*

[표 1]은 구문해석 결과의 유형을 제시한 것이다. CASE1 ~ 4 가 해석결과에 오류가 있는 경우로서 구문지식에 수정이 가해져야 한다. [표 2]는 해석결과에 오류가 있을 경우(CASE1 ~ 4) 원인 파악과 그에 따른 조치방법을 제시한다.

지금부터는 [표 2]의 내용을 좀더 살펴 본다. 한 문장의 구문해석 결과에 영향을 미치는 것이 사전정보와 문법, 두 가지라고 가정한다.

[표 2] 해석 결과 오류의 원인과 그 조치 방법

CASE \ 원인/조치	오류의 원인		조치 방법
1	사전정보	Wrong	Update
	문법	Omit	Add
		Wrong	Update
		Rough	Sub-Categorize
2	사전정보	Wrong	Update
	문법	Omit	Add
		Wrong	Update
		Rough	Sub-Categorize
3	사전정보	미등록어	어휘등록
		Wrong	Update
	문법	Omit	Add
		Wrong	Update
		Specific	Generalize
4	사전정보	Wrong	Update
	문법	Wrong, Omit	Update, Add
		Wrong, Wrong	Update, Update

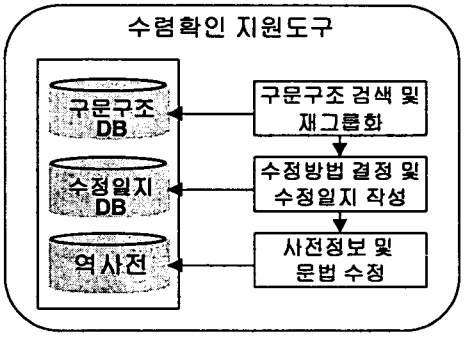
사전정보의 'Wrong' 표시에는 그 어휘의 정보 종류가 잘못된 경우와 단순한 타이핑 오류가 포함된다. 문법에는 'Omit, Wrong, Rough, Specific'의 4 가지 경우가 있다. 'Omit'은 해당되는 문법이 없어서 해석결과에 오류를 내는 경우이고, 'Wrong'은 단순한 타이핑 오류나 잘못 표현한 경우가 해당된다. 'Rough'는 기존 문법에 좀더 세분화가 필요한 경우로서 CASE3 을 제외한 나머지 경우에 해당되고 'Specific'은 기존 문법을 좀더 일반화시켜야 하는 경우로서 CASE3 에만 해당된다. CASE3 은 정문인데도 해석에 실패하는 경우인데 이는 기존 문법이 너무 세분화되어 있어서 좀더 일반화된 문법이 필요한 경우이다. CASE4 의 경우는 틀리게 지적한 부분과 원래 지적되었어야 했는데 지적되지 못한 부분 모두에 오류가 있는 경우로서 구문지식 수정 시에 함께 고려되어야 한다.

이 두 표는 사람이 개입되는 작업과정에서 사람의 판단영역을 좁히고 일관성을 유지시켜 줌으로써 작업과정의 실수를 줄이고 실수를 했다 하더라도 쉽게 복구할 수 있게 한다.

3.2 수렴 확인

구문해석 결과에 오류가 있을 때 작업자는 위에서 설명된 [표 1], [표 2]의 판단 기준에 따라 결과의 유형을 판단하고 그 원인을 파악하여 적절한 조치방법을 결정한다. 이 때 변화된 구문지식이 이전 단계까지 수용했던 언어현상도 함께 수용할 수 있어야 한다.

본 논문에서 제안하는 수렴성 확인 과정은 [그림 2]와 같다. [그림 2]의 과정은 지원도구를 이용하여 수행됨을 가정한다.



[그림 2] 수렴 확인 과정의 흐름도

구문구조 DB 는 이전 단계까지 분석된 문장의 구문구조를 저장한 것인데 저장 단위는 문법현상이 하나씩만 포함된 구문구조이다. 수정일지 DB 는 작업자가 인터페이스를 통해 직접 작성하는 것으로 사전정보나 문법 수정에 관하여 정해진 필드에 맞게 기록하는 것이다. 역사전(Inverted Dictionary)은 정보검색의 역화일의 개념을 도입한 것으로서[Salton89] 일반 사전과는 그 구조가 뒤바뀐 것이다. 즉 키(key)가 사전정보이고 그 데이터는 그 사전정보를 가진 어휘 리스트가 된다. 각각의 구조는 뒤의 예에서 제시한다.

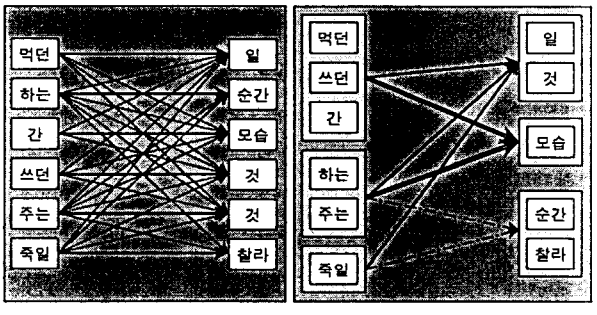
지금부터는 수렴 확인 과정을 설명한다. 먼저 현단계에서 수정하고자 하는 구문구조와 같은 것을 구문구조 DB 로부터 검색하여 수정 사항을 적용시킨 후 그 공기관계를 재그룹한다. 이 과정은 현단계에서의 수정 사항이 이전 단계에 영향을 주는지에 대한 여부를 확인하기 위해서이다. 다음은 재그룹한 결과에 따라 사전정보나 문법에 대한 수정 방법을 결정하여 수정일지를 작성한다. 수정작업에 대한 기록은 작업자에게 있어서 전문가시스템의 지식베이스와 같은 역할을 할 수 있다. 즉 작업자가 판단하기 어려운 상황에서는 예제를 제공해 주고 작업자

가 뒤늦게 실수를 깨달았을 때엔 쉽게 복구할 수 있게 해 주며 서로 다른 작업자 간의 일관성을 유지시켜 줄 수 있다. 이 단계에서는 [표 1], [표 2]에 의한 초벌 결정으로부터 좀더 구체적으로 어떤 정보나 문법을 어떻게 수정할 것인지를 결정한다. 결정이 끝나면 역사전과 지원도구를 이용하여 해당 어휘의 사전정보를 수정하거나 해당 문법을 일괄 수정한다.

예를 들어, "나는 학교에 갈 후, 몸이 아팠다"라는 잘못된 문장이 해석에 성공한 경우 이를 수정하는 작업은 다음과 같다. 먼저 [표 1], [표 2]로부터 해석결과 유형이 CASE2(비문인데 해석에 성공)이고 '갈 후(etm|nc)'[1]에 적용되는 문법이 너무 포괄적이어서 세분화가 필요하다는 것까지 결정한다. 다음은 수정해야 하는 구문구조인 'etm|nc'를 [그림 3]과 같은 구조를 가진 구문구조 DB로부터 검색하여 [그림 4]와 같이 공기관계를 다시 설정하여 그룹화한다.

etm|nc	먹던|일
	하는|순간
	죽일|찰라
	쓰던|것

[그림 3] 구문구조 DB의 예



[그림 4] 공기관계에 따른 재그룹화

[그림 4]의 왼쪽은 재그룹하기 전의 상태인데 구문구조 DB로부터 검색된 것은 실선으로 연결된 구조이다. 점선은 작업자가 새로이 공기관계를 설정시켜 준 것이다. 이 작업은 지원도구를 이용함으로써 간단하게 수행될 수 있다. [그림 4]의 오른쪽은 재그룹한 후의 상태인데 같은 공기관계를 가지는 것끼리 묶은 것이다. 재그룹된 결과에 따라 구체적인 수정방식을 결정하여[2] 수정일지 DB에 기록한 후 해당 어휘의 사전정보나 문법을 수정한다. 다음은 수정일지를 작성한 예이다.

CASE 번호	2			
원문의 위치	문서번호	100	문장번호	53
해석결과	수정전	(갈 O 후)		
	수정후	(갈 X 후)		
사전정보	수정전	ㄹ	SCAT|etm	
		후	SCAT|nc	
	수정후	ㄹ	SCAT|etm*etm|etm2	
		후	SCAT|nc*etm|+etm3	
문법	수정전			
	수정후			
조치 사항	(관형형어미 O 명사) 세분화 관형형어미->'는', 'ㄹ(을)', 'ㄴ(은), 던'->사전정보 수정 명사->관형형어미와의 공기관계를 세부자질화->사전정보 수정			

[그림 5] 수정일지의 예

[그림 5]에 기록된 변경사항을 살펴 보면, 'ㄹ'이 'etm'이라는 카테고리에서 새로운 서브카테고리인 'etm2'를 가지게 되었고 '후'는 'etm' 중 'ㄴ(은),던'과 같은 'etm3'와 공기할 수 있으므로 'etm|+etm3'라는 새로운 자질-값을 가지게 되었다.

3.3 문제점 및 고찰

기존 연구를 살펴 보면, 검증되지 않은 태그 세트로 태깅된 말뭉치로부터 통계정보를 추출하여 문법이나 구문구조를 자동으로 획득하는 방식이 활발히 연구되고 있는데[Shih95][Wilms95][Margerman90] 이 방식이 속도면에서는 빠를 수 있으나 그 정확도는 보장할 수 없다. 태그 세트의 품질에 따라 통계정보의 품질 또한 결정되므로 충분한 검증이 이루어지지 않은 태그 세트로는 의미있는 통계정보의 추출이 어렵기 때문이다.

제안된 방식으로 언어지식을 구축하다 보면 기존 방식에 비해 속도면에서 늦어질 수 있다. 그러나 많은 양을 빨리 구축하였다고 해도 효과적으로 활용되지 않으면 아

[1] 본 논문에서 적용한 구문 카테고리로서 etm은 관형형어미를, nc는 보통명사를 가리킨다.

[2] 각 그룹에 새로운 카테고리를 부여하는데, 새로운 서브 카테고리로 삼을 것인지, 기존 카테고리의 새로운 자질-값으로 표현할 것인지 등을 결정한다.

무 소용이 없게 되므로 누구나 믿고 활용할 수 있는 고품질의 언어지식 구축을 위하여 제안된 방식을 적용할 것을 주장한다. 또한 제안된 방식을 지원 도구 개발을 통해 자동화한다면 정확하면서도 신속한 정보 구축 작업이 이루어질 수 있을 것이다. 본 연구에서는 지원도구의 일부가 개발되어 있고 앞으로 완성하여 적극 활용할 계획이다.

4 결론

본 논문에서는 지금까지 언어지식의 질적 수렴이 보장되지 못한 원인을 분석하고 언어지식의 수렴을 보장할 수 있는 방법을 연구하였다. 지금까지 양적 확장에 많이 치중하여 구축된 언어지식이 구축에 들인 노력이나 비용만큼 가치를 발휘하지 못하고 있는 것은 양적 확장과 동시에 질적 확장이 이루어지지 못했다는 것을 알려준다.

언어지식 구축 작업은 완전 자동화되기는 힘들지만 그 과정을 최대한 자동화, 공정화하여야 한다. 사람이 개입되는 작업은 사람마다 서로의 판단기준이나 지식의 정도가 다르고 같은 사람이라도 시간이 지남에 따라 판단기준이 달라지거나 실수를 할 수 있는데 작업 과정의 자동화, 공정화는 이러한 문제를 극복해 줄 수 있기 때문이다. 이를 위하여 본 논문에서는 전반적인 구축 과정 중에서 특히 사람의 판단이 주가 되는 과정인 구문해석 결과의 분석을 통하여 오류를 복구하는 작업 과정을 자동화, 공정화하는데에 중점을 두었다. 특히 이 과정에서 언어지식의 변화가 이상적인 언어지식으로 수렴해 나가고 있는지를 확인할 수 있는 방법을 제안하였다.

본 논문에서 제안한 방법이 아직 완전히 증명이 되진 않았지만 언어지식 구축 작업이 더 이상은 노동 집중식(labor intensive) 작업이 아닌 체계적이고도 자동화된 방법에 의하여 수행되어야 하고 양적 확장 위주에서 질적 확장에 중점을 두어야 한다는 것은 분명하다. 앞으로 본 논문에서 제안한 방법으로 언어지식을 구축하면서 질적 수렴을 가장 잘 보장할 수 있는 방법으로 개선해 나갈 것이며 동시에 지원 도구 개발을 해나갈 것이다.

5 참고문헌

[시스 97] 시스템공학연구소, *STEP2000 제3차년도 최종보고서*, 1997

[이승 94] 이승선, 송주원, 황규영, 최기선, "TRIE 구조를 이용한 한국어 전자 사전을 위한 데이터베이스 인덱스 구조", 한국정보과학회 봄 학술발표논문집, Vol. 21, No. 1, pp. 849-852, 1994

[이현 96] 이현아, 박재득, 장명길, 박수준, 박동인, "구문적 언어지식 획득 과정의 문제점 분석 및 지원도구설계", 제 8 회 한글 및 한국어정보처리 학술발표논문집, pp. 489-496, 1996

[Allen95] James Allen, *Natural Language Understanding*, The Benjamin/Cummings Publishing Company, 2nd Edition, 1995

[Lee97] Hyun-A Lee, Jae-Deuk Park, Soojun Park, Myung-Gil Jang, Dong-In Park, "The Problem Analyses on Syntactic Knowledge Acquisition and The Design of Supporting Tools", Proceedings Of the 17th International Conference on Computer Processing of Oriental Languages, Hong Kong, pp. 77 – 82, 1997

[Margerman90] D. M. Margerman, M. P. Marcus, "Parsing a Natural Language Using Mutual Information Statistics", Proceedings of AAAI, 1990

[Salton89] Gerald Salton, *Automatic Text Processing*, Addison-Wesley, 1989

[Shih95] H-H. Shih, S. J. Young and N. P. Waegner, "An inference approach to grammar construction", Computer Speech and Language, No. 9, pp. 235-256, 1995

[Wilms95] Geert Jan Wilms, "Automated induction of a lexical sublanguage grammar using a hybrid system of corpus and knowledge-based techniques", A dissertation submitted to the faculty of Mississippi State University, 1995