구문적 언어지식 획득 과정의 문제점 분석 및 지원도구 설계

이 현아, 박 재득, 장명길, 박수준, 박동인

시스템공학연구소 자연어 정보처리 연구부

Problem Analysis on Syntactic Linguistic Knowledge Acquisition and Design of a Supporting Tool

Hyun-A Lee, Jae Deuk Park, Myung-Gil Jang, Soojun Park, Dong-In Park

Dept. of Natural Language Information Processing, SERI

요 약 문

자연어 처리에서 언어에 대한 지식은 전자사전과 문법규칙으로 구성되어 서로 상보적 관계에 있고, 각 어휘에 대한 품사 및 기타 자질-값에 의해 매개된다. 이러한 언어지식을 전통적인 방법에서는 국어자료의 분석에 경험이 많은 언어전문가의 직관에 다분히 의존하여 정의하였고, 말뭉치를 이용한 자동 획득 기법에서는 태그세트를 먼저 설정하고, 이 태그를 원시 말뭉치에 부착하여 태깅된 말뭉치로부터 자동으로 통계적 분석을 통하여 획득한다. 그런데 두가지 접근방법이 가지고 있는 공통적인 문제점은 품사나 자질-값의 정의 및 할당기준, 선악의 평가기준, 튜닝에 대한 적극적 대처 등이 마련되어 있지 않다는 점이다. 이 연구에서는 이러한 문제점의 발생원인을 말뭉치 분석 과정에서 살펴보고, 품사 및 자질-값의 설정과 할당기준을 마련하는 방법론 및 이를 적극적으로 지원하는 도구를 설계한다.

1. 서론

자연어 처리 연구에서 파라다임이나 기법 상의 변화는 많이 있었지만 언어에 대한 지식은 보통 전자사전과 규칙의 형태로 표현되어 왔다. 어휘에 대한 개별적 성질을 담고 있는 사전과 어휘 범주 간의 일반적 성질을 반영하는 규칙의 충실도는 언어지식의 망라성의 측도이며, 시스템의 품질을 좌우하는 주요 요소로 여전히 남아 있다. 그리고, 전자사전 및 문법규칙은 서로 상호보완적 관계에 있으며, 각 어휘에 대한 품사 및 자질-값(feature-value)에 의해 매개된다.

이러한 언어지식을 구축하는 데는 크게 규칙 기반 방법과 말뭉치 기반 방법의 두가지가 있다. 규칙 기반 방법은, 선험적 지식을 이용하여 품사 및 자질-값 체계를 미리 설정한 후 사전과 문법을 그에 따라 구축한다[3]. 물론 선험적 지식은 여러 전문가의 검증을 거친 지식으로서 어느 정도의 정확성을 보유하고 있지만 새로운 현상에 대한 대처능력이 부족하다. 규칙 기반 방법을 보완하기 위해 시도되는 말뭉치 기반 방법은, 말뭉치로부터 추출되는 용례정보나 통계정보 등을 이용하여 언어지식을 획득한다[1][2]. 최근 말뭉치로부터 문법자동획득[5][6], 구문구조 자동획득[9], 용언하위범주화 자동획득[8], 품사분류 자동획득[4][7] 등에 관한 연구가 활발히 진행되고 있다. 그런데 이들 연구는 문법과 사전, 품사태그분류 각각을 독립적으로 획득하는 접근방법이다.

그러나 그림 1 에서 나타내려고 한 바와 같이 구문규칙이나 사전의 내용은 품사태그 및 자질-값의 종류에 좌우되고 품사태그는 구문규칙에 사용되기 위해 정의되므로 구문규칙, 사전, 품사태그, 자질-값은 서로 연관되어 있어 각각을 독립적으로 획득하고자 하는 방법은 부적합하다. 예를 들어, 문법을 독립적으로 획득하기 위하여 태그세트를 미리 고정하게 되면, 말뭉치의 성질을 미리 규정하는 것이기 때문에 가능한 규칙과 사전 내용의 범위도 한정되고 만다. 따라서 실제 말뭉치에 나타나는 정문 및 비문들에 대한 변별력과 구문구조 분석의 정도가 떨어질 것이므로 정도를 높이기 위한 튜닝을 필연적으로 거쳐야 한다. 따라서 태그세트 자체도 정문과 비

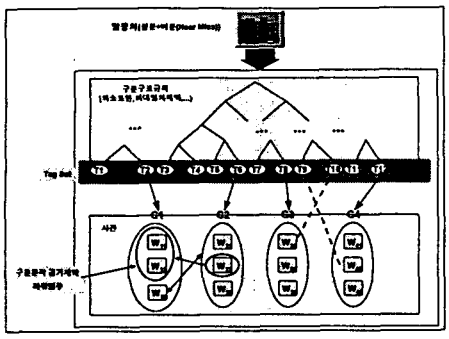


그림 1 태그세트, 문법, 자질-값, 사전, 말뭉치의 관계

문이 함께 포함된 말뭉치의 분석과 동시에 문법, 자질-값 세트, 사전까지 고려하여 도출되어야 한다. 즉, 이들 언어지식은 동시에 상호작용적이며 병발적으로 구성되어야 하며 그 과정은 문장단위 또는 새로운 현상단위로써 점진적으로 이루어져야 한다는 것이 본 논문의 주장이다.

이를 위해 본 논문에서는 품사 및 자질-값의 설정과 할당기준을 마련하는 방법론과 이를 적극적으로 지원하는 도구인 언어지식 개발 환경(Linguistic Knowledge Development Environment, LKDE)의 설계 및 구현을 목표로 한다. LKDE는 정문과 비문이 포함된 테스트 문장들의 해석 결과를 품사 및 자질-값 체계와 문법의 구축에 즉각 반영되게 하고, 바뀌어진 품사 및 자질-값 체계와 문법이 기관찰된 현상을 커버할 수 있는지 검사하여 다음 테스트 문장의 해석에 즉각 적용시켜 언어지식을 점진적으로 정비해 나가며 각각의 과정을 지원할 수 있는 도구를 포함한다. LKDE는 직관적인 방법에 대한 객관적 보완 방법으로 제시하는 것이며 과다한 작업량으로 일관했던 언어지식 구축 과정을 반자동화하여 시간과 노력을 절약할 수 있게 한다.

2. 기존의 언어지식 구축 방법의 문제점 분석

사전은 용도나 사용기법에 따라서 여러가지 형태가 있을 수 있으나 자연어 처리용 전자사전은 보통 자질-값의 리스트나 구조로 표현된다. 품사 및 자질-값은 한 단어의 카테고리 이름과 해당 단어가 문장 내에서 가질 수 있는 제약관계를 나타낸다. 구문적 현상의 예를 들면, "어느"를 표현하는 품사 즉 카테고리 이름은 "관형사"이고, "명사만을 수식한다"는 제약관계가 자질-값으로 나타나야 한다.

언어처리는 여러 레벨의 처리를 거쳐야 하고 따라서 여러 종류의 정보가 필요하나 본 논문에서는 구문적 현상만을 고려한다. 그리고, 자질-값 중에서는 해석기나 생성기에서 직접 쓰이는 정보에 대해서만 다루기로 한다.

2.1. 규칙기반 방법에서의 문제점

규칙기반 자연어 처리 연구 개발자들은 주로 해석이나 생성 알고리즘과 문법 표현, 포말리즘이나 사전구조 등의 효율적 표현 및 처리에 치중하고, 문법 규칙이나 사전에 등록될 정보의 종류를 망라적으로 정확히 정의하는 것과 대규모 사전 및 문법 규칙 작성에는 상대적으로 적은 비중을 두어 왔다. 그 이유는 주로 대규모 사전 및 문법 규칙의 작성에 소요되는 시간과 노력 때문에 아예 엄두를 못내거나 중요성은 알지만 언어학자나 국어학자들이 전담할 일이라고 넘겨왔기 때문이다. 그러나 언어학자나 국어학자들도 나름대로의 이론 제시에 치중하여 이와 같은 일에는 관심을 기울이지 않는 경향이 있으며, 그렇지 않다 하더라도 아래에서 열거되는 몇 가지 문제점이 있다.

- 미리 어느 정도의 국어학 분야의 선행 연구 후에 시한과 조사대상을 한정시켜 놓고 품사 및 자질-값을 고정시키고 이에 의존해서 나중에 값을 지정한다.
- 미리 작성된 사전 작성 메뉴얼의 범위 내에서 모든 단어의 품사 및 자질-값을 할당한다.

- 품사 및 자질-값 할당 기준을 적용할 때 어떤 것을 선택해야 할지와 새로 추가를 해야 하는가에 대한 명확한 판단기준이 없으며 일반적인 방법론도 제시되지 않는다.
- 지나친 세분류나 부족한 분류로 인해 나중에 별 필요가 없을 분류를 미리 한다. 결국 불필요한 사전 입력 작업이 가중된다.
- 오류 발생 시에 적절하고 신속하게 반영하는 적극적인 방법론이 없다.

2.2. 통계적 방법에서의 문제점

최근 규칙기반 방법의 단점을 보완하기 위해 통계적 방법이 시도되고 있다. 통계적 방법에서는 먼저 일정한 품사 태그 세트를 정하고 훈련 말뭉치를 정해진 품사 태그 세트로 수동 태깅하여 구축한 후 통계정보를 추출한다. 결국 품사 태그 세트의 질에 따라 추출되는 통계정보의 정확성이 좌우되는 것이다. 그러나 기존의 통계적 방법에서는 주어진 품사 태그 세트에 대한 평가보다는 그것은 맞다고 가정하고 그 후의 통계정보 추출에 치중하고 있다.

2.3. 정리

두 방법에서 공통적으로 나타나는 문제점을 살펴보면, 바로 문법규칙이나 품사 태그, 자질-값의 설정 방법이 잘 정의되어 있지 않다는 것이다. 지금까지 어떤 정해진 원칙이나 원리에 의해 접근하지 않고 직관적인 관행에 따라 해온 것이다.

이러한 문제점을 극복하기 위해 본 논문에서는 문법규칙이나 품사태그, 자질-값 등의 언어지식(linguistic knowledge) 구축하는 데 있어서 품사 및 자질-값의 할당 기준과 튜닝에 대한 적극적 대처 등이 잘 정의된 언어지식 개발 환경과 그에 대한 평가기준을 제시하고자 한다.

3. 언어지식 개발 환경(Linguistic Knowledge Development Environment)

본 논문에서는 실제 문장의 해석이나 생성을 위해 꼭 필요한 정보를 제공할 수 있는 언어지식의 구축을 위하여 언어지식의 점진적 수렴성을 추구하는 새로운 언어지식 개발 환경을 제안한다. 즉, 언어지식이 실제로 문장의 해석이나 생성을 위하여 사용되어짐에도 불구하고 해석기나 생성기와는 별개로 구축되어져 온 문제점을 개선하고 언어지식의 구축과 튜닝에 드는 많은 시간과 노력을 줄일 수 있도록 해석기와 여러 지원 도구가 맞물린 개발환경을 제안하는 것이다. 본 논문에서 제안하는 언어지식 개발 환경(이하

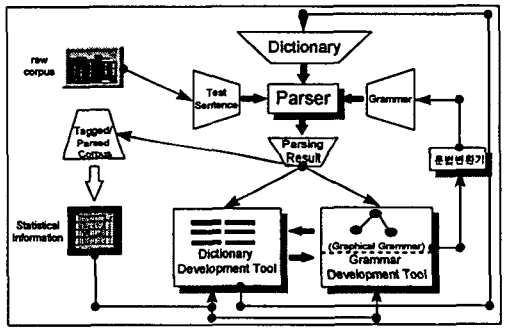


그림 2 LKDE의 전체 구성도

LKDE)의 전체 구성도는 그림 2와 같다. LKDE는 크게 해석기, 사전 개발 도구, 문법(규칙) 개발 도구, 문법 변환기로 구성되어 있다.

정문과 비문이 함께 포함된 테스트 문장을 해석기로 해석한 후 결과(구문 트리)를 표 1에 제시된 경우 중 어디에 속하는 지를 사람이 판단하여 표 2에 제시되어 있는 적절한 액션을 취하는 감시자 모드(supervised mode)에서 문법과 하위범주 정보를 튜닝해 나간다.

표 1은 해석기를 통한 문장 해석결과의 8가지 유형을 제시해 놓은 것이다. 이 중 오류가 생기는 경우는 4가지이다. 정문을 바르게 accept하는 정해석과 비문을 바르게 reject하는 정지적

표 1 문장 해석 결과의 유형

		정문	비문
accept	정해석	PASS	NULL
	오해석	경우 1	경우 2
reject	정지적	NULL	PASS
	오지적	경우 3	경우 4

표 2 해석결과 오류의 원인과 처치방법

원인 1/2		처치 1/2
사전 자질-값 정확	규칙결여	규칙 첨가
	규칙오류	규칙 수정
사전 자질-값 부정확	규칙결여	사전 자질-값 수정->규칙첨가
	규칙오류	사전 자질-값 수정->규칙수정
원인 3/4		처치 3/4
사전 항목 결여		사전 항목 추가
사전 자질-값 정확	규칙결여	규칙 첨가
	규칙오류	규칙 수정
사전 자질-값 부정확	규칙결여	사전 자질-값 수정->규칙첨가
	규칙오류	사전 자질-값 수정->규칙수정

만이 옳은 경우이다. 정문의 경우, 틀리게 accept 하는 오해석과 reject 하는 오지적이 오류이고 비문의 경우, accept 하는 오해석과 reject 하되 지적한 부분이 틀린 오지적이 오류이다. 표 2 에 제시된 4 가지 오류의 원인과 그 처치방법을 살펴보면 사전의 자질-값이 정확한 경우, 문장에 나타난 현상을 다룰 수 있는 문법규칙의 결여나 오기술(wrong description)이 원인이고 그 처치방법은 적절한 규칙의 첨가나 수정이다. 만약 사전의 자질-값이 부정확하다면 정확한 정보를 입력시키고 그에 따른 규칙의 첨가나 수정을 해야 한다. 사전 항목이 결여된 경우는 오지적에만 해당되는 원인이다. 사전 항목이 결여되어 있으면 정문이든 비문이든 accept 하지 못하기 때문이다.

이렇듯 문장 해석 결과에서 발견되는 오류를 통해 사전의 자질-값이나 문법규칙을 수정, 보완함으로써 점차 오류를 적게 발생시키는 언어지식으로 정비되어 나갈 수 있다.

이후에는 문장의 해석결과로부터 사전 자질-값과 문법규칙의 튜닝을 적극적으로 지원하는 LKDE 을 각 구성요소 별로 설명하기로 한다.

3.1. 해석기(parser)

정문과 비문이 함께 포함된 테스트 문장을 해석하여 결과(구문 트리)의 유형이 표 1 에서 제시된 경우들 중 어디에 해당되는 지를 사람이 판단하고, 그에 해당하는 적절한 액션을 취함으로써 문법과 하위범주 정보의 정비가 이루어지게 한다. 이를 위해 LKDE 의 해석기가 가져야 할 기능은 다음과 같다.

- 해석결과를 트리로 표현한다.
- 튜닝된 문법의 즉각적인 적용을 위해 파싱 엔진과 문법 해석 엔진이 독립적으로 존재해야 한다.

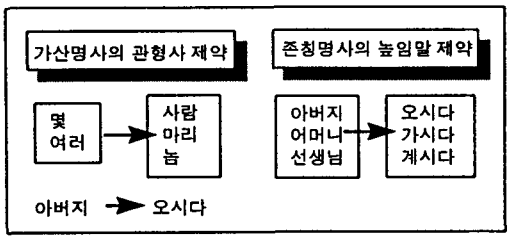
해석결과를 트리로 표현하면 정지적과 오지적 여부에 대한 사람의 판단을 용이하게 한다. 또한 파싱 엔진과 문법 해석 엔진이 독립적으로 존재하면 문법의 변화가 바로 파싱에 적용될 수 있어 튜닝에 드는 시간을 그만큼 줄일 수 있는 효과를 가져온다.

3.2. 사전 개발 도구(Dictionary Development Tool)

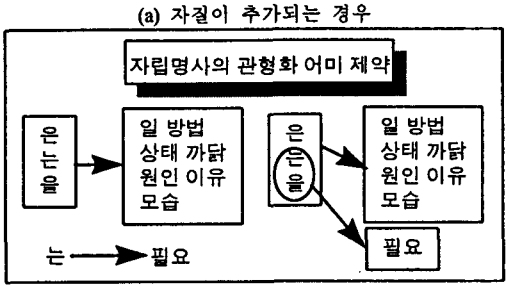
어휘에 대한 정보를 사전에 입력할 때 요구되는 기능을 열거하면 다음과 같다.

① 한 어휘에 어떤 정보를 입력할 것인지에 대한 입력자의 판단이 용이해야 한다.
② 입력된 정보의 수정이 용이해야 한다.
③ 해석기로의 적용이 용이해야 한다. 즉 따로 사전을 컴파일하지 않고 바로 적용할 수 있어야 한다.

이러한 요구들을 만족시키기 위해 본 논문에서는 사전 개발 도구를 설계 및 구현하였다. 먼저 첫번째 기능을 위해 사전 정보 체계를 계층적으로 구성하여 정보 입력자에게 상위계층에



(a) 자질이 추가되는 경우



(b) 값이 추가되는 경우

그림 3 자질-값의 변화가 생기는 현상

서 하위계층의 순서로 정보를 제시함으로써 일관성 있는 판단을 가능하게 한다. 두번째 기능을 위해 수정하고자 하는 단어의 정보를 바로 액세스하고 수정한 후에 바로 저장할 수 있도록 하였다. 세번째 기능을 위해 사전의 저장구조로서 gdbm 데이타베이스 시스템을 사용하여 정보가 입력, 저장되면 따로 사전을 컴파일하지 않고도 바로 해석기로 적용될 수 있도록 하였다.

3.3. 문법 개발 도구(Grammar Development Tool)와 언어지식의 점진적 수렴

선언적으로 표현한 종래의 문법은 해석기 등에 적용되기는 용이하지만 사람이 문법을 이해하기에는 부적당하다. 본 논문에서는 사람이 이해하기 쉬운 문법 표현으로서 그래프 표현방식을 제안한다. 문법은 기본적으로 자질-값(feature-value) 형태를 취하는데 테스트 문장의 파싱결과로부터 새로운 현상이 발견될 때마다 자질-값의 추가가 일어나서 현상에 대한 커버력이 점점 강화되어야 한다. 이에 관한 예를 그림 3 에서 들었다.

그림 3(a)에서 명사가 일부 관형사를 제약하는 자질에 의해 셀 수 있는 것과 없는 것으로 명사를 분류했으나 새로운 현상인 '아버지가 오시다'를 처리하기 위해서 높임 용언에 의해 명사를 제약하는 자질이 추가되었다. 그림 3(b)에서는 자립명사의 관형화 어미 제약 자질 내에 새로운 현상인 '~는 필요'를 처리하기 위해 관형화 어미 중 '는,을'과만 결합가능하다는 값이 추가되었다. '~는 필요'가 기존에 밝혀진 자립명사의 관형화 어미 제약관계에 포함되기 위해서는 '~은 필요'와 '~을 필요'가 모두 가능한지를 검사해야 한다. 이 단계에서 불가능한 관계가 발견되면 단순한 '자립명사의 관형화 어미 제약관계'에서 좀더 세분되어야 할 것이다[그림 3(b)].

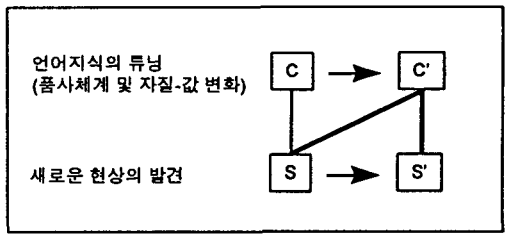


그림 4 언어지식의 점진적 수렴

즉, 어떤 새로운 현상(S')이 발견되면 기존의 분류체계나 자질-값(C)으로 완전히 커버되는지를 검사하여 커버가 되지 않는다면 기존의 분류체계 및 자질-값을 새로운 현상을 커버할 수 있도록 변화시켜야 하고 동시에 기관찰된 현상(S)에 대해서 변화된 분류체계 및 자질-값(C')에 의해 완전한 커버가 되는지를 확인해야만 점진적인 언어지식의 수렴이 이루어질 수 있다[그림 4]. 이 때, 변화된 분류체계 및 자질-값을 기관찰된 현상의 어휘들에 재할당해야 하는데 이에 관한 정형화된 방법론의 정립은 향후 과제로 연구할 것이다.

본 논문의 문법 개발 도구는 자질-값의 추가나 세분류를 용이하게 하기 위해 초기 그래프로부터 새로운 문법이 도출되거나 기존 문법을 수정할 필요가 있을 때에는 그래프의 노드와 에

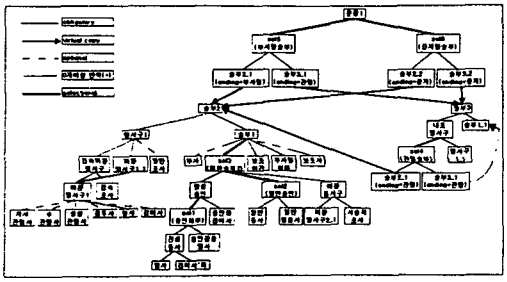
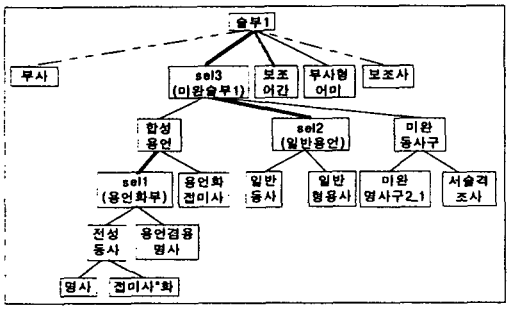
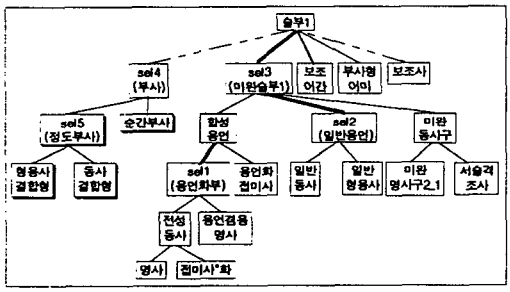


그림 5 문법의 그래프 표현의 예



(a)



(b)

그림 6 문법규칙 수정의 예

지를 추가하거나 에지의 연결상태를 바꾸는 등의 방식을 취한다.

그림 5는 선험적 지식에 기반한 문법들을 문법 개발 도구를 이용하여 그래프로 표현한 결과이다. 그림 5의 그래프에는 여러 종류의 에지가 존재한다. 이는 한국어 문장에서의 생략 현상이나 부분 자유어순 현상 등이 반영되었기 때문이다.

그림 6과 그림 7에서는 사전의 자질-값과 문법규칙을 첨가 또는 수정하는 방법과 도구의

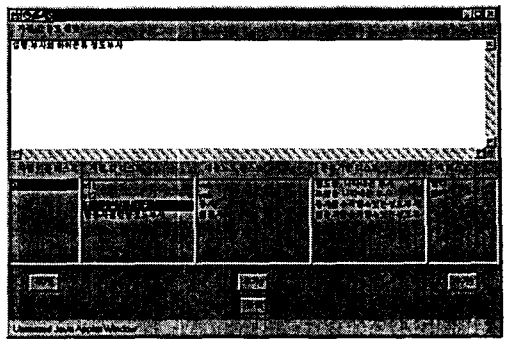


그림 7 자질-값 수정 지원 도구

설계를 보인다. 그림 6(a)에서 술부의 구성요소(constituency) 중의 하나로 "부사"가 제시되었는데 이는 모든 부사와 모든 용언이 술부를 이룰 수 있다는 것을 뜻한다. 그러나 테스트 문장 중 "영희가 매우 달린다."라는 비문이 accept 되므로 "부사"가 그림 6(b)와 같이 세분류되어야 하고 동시에 그림 7의 도구를 이용하여 "부사"의 하위분류 정보를 입력한다. 이 때, 분류기준을 기록해 놓아 다른 어휘에 자질-값을 할당할 때 객관적인 기준으로 삼는다.

3.4. 문법 변환기

새로운 자질-값의 추가나 세분류가 일어난 문법을 다른 테스트 문장의 해석에 즉각 적용시키기 위해 그래프로 표현된 문법을 해석기에서 쓸 수 있는 형태로 변환해야 한다. 본 논문에서는 이를 위해 그래프로 표현된 문법을 선언적 형태로 변환시키는 문법 변환기를 설계하였다. 앞서 언급하였듯이, 문법은 해석기 엔진과는 독립적으로 존재하게 하여 문법의 변화가 즉각 해석기의 해석 수행에 적용될 수 있게 하여야 한다. 문법의 변화로 인해 해석기 전체를 다시 컴파일하는 것과 같은 부가 작업은 튜닝 작업을 그만큼 더디게 하는 요인이 된다.

4. 언어지식의 평가 기준

먼저 언어지식표현기법(사전표현양식, 문법기술

언어, 품사태그세트)은 수정이 용이해야 하고(flexibility) 사람이 보았을 때 기호체계가 어떤 의미인지 이해하기 쉬어야 한다(readability, interpretability, expressive power). 그러나 이러한 기준으로 언어지식표현기법을 정하고 같은 목적 아래 같은 말뭉치와 같은 해석기를 사용하고서도 언어지식전문가(문법/사전 전문가)들은 서로 다른 언어지식을 구축할 수 있다. 따라서 기존의 파싱 성공률 외에 이들이 구축한 언어지식을 평가할 수 있는 기준이 필요하다.

본 논문에서는 주어진 말뭉치에 대하여 획득된 언어지식을 평가하는 기준으로, 관찰 데이타와의 일치도(maximal goodness of fit)와 표현지식의 효율성(minimal representation)을 제시한다. 이 두 기준은 어느 한쪽을 높이려면 다른 한쪽이 낮아지는 trade-off 관계에 있고 따라서 주어진 말뭉치에 대하여 최대 적합도와 최소 표현기술(description)을 동시에 만족시킬 수 있는 것이 최적의 언어지식이다.

5. 결론

본 논문의 연구는 언어정보베이스의 핵심을 이루는 전자사전의 자질-값의 도출과 각 어휘에 대한 자질-값의 할당 기준의 도출, 그리고 이러한 사전과 상호보완적으로 관찰된 언어현상을 망라하는 문법규칙의 도출 과정을 분석하고 이에 부합하는 방법론을 정립하는 것이다. 이러한 도출 과정이 그동안 언어학자나 국어문법학자 혹은 전산언어학자들의 언어 직관적인 인지과정으로서 간섭이나 의문을 품을 수 없는 신성불가침의 영역으로 생각되어 왔으나 본 논문에서는 이러한 과정을 객관화하여 보다 정교한 언어학적 지식 획득을 지원하려는 시도를 하였다.

본 논문에서 제시한 언어지식 구축 기법은 개발 초기의 정도 높은 핵심적 기초 언어지식을 구축하는 데에 기여한다. 본 논문에서 제시한 방법으로 개발된 핵심적 기초 언어지식은 통계적 방법에서의 훈련 말뭉치 구축에 쓰이는 태그세트로 제공되어 보다 정확한 통계정보를 추출하게 하고 해석기의 지식베이스로 제공되어서도 오류 발생이 작을 것이다.

본 논문에서 제안한 언어지식 개발 환경이 효과적이기 위해서는 정문과 비문이 함께 포함된 말뭉치가 절대적으로 필요함에도 기존의 말뭉치에는 비문이 거의 나타나지 않는 문제점이 있다. 그리고 본 논문에서 제안한 언어지식 개발 환경에서 새로운 현상을 처리해 나가면 언어지식이 점진적으로 수렴해 나가느냐 하는 문제이다. 이의 증명은 실제 언어지식의 구축과 말뭉치 분석을 통해 이루어질 것이다. 그러나 본 논문에서는, 어느 정도의 프로토타입이 개발되면 그 다음의 새로운 현상이나 기존에 부여된 값의 오류 수정 등의 튜닝은 일관성있고 순서적으로 정해진 절차에 따라서 수행되어야 하는 인지과정의 하나로 보고 이를 지원하는 언어지식 개발 환경을 설계 및 일부 구현하였고 이를 이용하여 소량의 말뭉치에 대한 언어지식의 점진적 수렴성을 경험하였으며 대량의 말뭉치에 대해서도 같은 결과를 기대한다.

6. 참고 문헌

1. 이상섭, "뭉치 언어학: 사전 편찬의 필수적 개념", 제1회 한글 및 한국어 정보처리 학술발표논문집, pp. 73-76, 1989
2. 이호석, 김영택, "영한 변환사전 생성을 위한 말뭉치에 기반한 연어와 실용어의 자동 추출", 한국정보과학회 논문지 Vol. 21, No. 11, pp. 2110-2117, 1994
3. 서정수, "국어구문론 연구", 탑출판사, 1983
4. 이공주, 김재훈, 김길창, "한국어에서의 단어 자동 분류와 품사 분류 체계", 1996년도 한국정보과학회 봄

학술발표논문집, Vol. 23, No. 1, pp. 961-964, 1996

5. 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
6. 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
7. Hinrich Schutze, "Part-of-speech induction from scratch", 31st Annual Meeting of the Association for Computational Linguistics(ACL 93), pp. 251-258, 1993
8. Michael R. Brent and Robert C. Berwick, "Automatic acquisition of subcategorization frames from tagged text", Proceedings of Speech and Natural Language Workshop, Pacific Grove, California, February, pp. 342-345, 1991
9. D. M. Margerman, M. P. Marcus, "Parsing a Natural Language Using Mutual Information Statistics", Proceedings of AAAI, 1990