독영, 영한 자동번역 시스템 연결에 의한 인터넷 독한 자동번역

최승권(한국전자통신연구원)

1. 머리말

인터넷의 발달과 웹 브라우저의 보급으로 우리는 전 세계로부터 영어나 독일어로 된 정보를 더욱 편리하게 얻을 수 있게 되었다. 이와 관련하여 전 세계의 언어별 웹문서 분포를 살펴보면 다음과 같다:

【도표 1】 언어별 전 세계 웹사이트 분포(1998)

순위	언어	웹페이지 수	퍼센트(%)
1	영어	2722	84.0
2	독일어	147	4.5
3	일본어	101	3.1
4	불어	59	1.8
5	스페인어	38	1.2
6	스웨덴어	35	1.1
7	이태리어	31	1.0
8	포르투갈어	21	0.7
9	네덜란드어	20	0.6
10	노드웨이어	19	0.6
11	핀란드어	14	0.4
12	체코어	11	0.3
13	덴마크어	9	0.3
14	러시아어	8	0.3
15	말레이시아어	4	0.1
전체		3239	100

도표[1]은 임의로 추출된 3239개의 웹문서 중에서 언어별 웹문서의 분포를 살펴본 것으로 인터넷의 대부분 정보가 영어로 이루어져 있지만 영어 이외에 기타의 언어에 의한 인터넷에서의 정보 제공율도 16%나 차지한다는 것을 알 수 있다.

1.1. 인터넷에서 독일어 번역의 필요성

도표[1]로부터 알 수 있듯이 인터넷에서 독일어가 차지하는 비중은 그리 크지 않다. 하지만 인터넷에서 독일어가 차지하는 비중이 작다고 하더라도 본 저자는 다음과 같은 이유에서 독일어 자동번역의 필요성을 강조하고자 한다.

- ▶ 전문정보의 획득 필요성: 독일 연방경제부의 『Info 2000: Deutschlands Weg in die Informationsgesellschaft』에 따르면 향후 독일의 학술적 기술적 정보는 인쇄된 문서보다는 전자화된 문서로 대용량 구축될 계획으로 대용량의 전자화된 독일 정보를 획득할 필요성이 있다.
- ▶ 일반인에 대한 독일어 정보 제공 필요성: 독일어에 대한 지식이 없는 사람들에게 독일어 웹문서의 독일어 정보를 한국어로 번역하여 제공할 필요성이 있다.
- ▶ 전공자에 대한 대용량 독일어 초벌번역 제공 필요성: 대용량의 독일어 웹문서를 독일어 전공자라도 매일같이 충분히 소화할 수 없기 때문에 초벌로 자동번역된 대용량의 웹문서를 독일어 전공자에게도 제공할 필요가 있다.

1.2. 본 논문의 구성

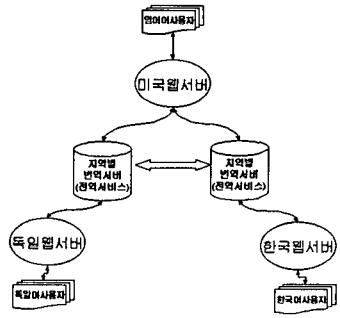
본 논문에서는 독일어를 원시언어 Quellsprache로 한 한국어로의 자동번역 방법에 대해 기술하고자 한다. 본 논문의 구성은 다음과 같다.

2장에서는 본 논문에서 논의하고자 하는 인터넷 독한 자동번역 방법과 이 방법을 구성하고 있는 독영 자동번역 방법과 영한 자동번역 방법에 대해 개괄적으로 기술하고자 한다. 3장에서는 대조 언어학적 관점에서의 독한 자동번역의 문제를 언급하고 그에 대한 본 논문에서의 해결방법을 기술하고자 한다. 4장에서는 더욱 좋은 번역품질을 만들어 내기 위한 두 단계 전산문법에 대해 기술하고자 한다. 5장에서는 독한 자동번역에 사용된 영

한 자동번역의 번역지식(사전과 규칙)의 크기와 독한 자동번역의 실험결과를 제시할 것이다.

2. 인터넷 독한 자동번역의 방법

본 논문에서의 인터넷 독한 자동번역의 방법은 우선 한국어로 번역하고자 하는 독일어 웹문서를 영어로 자동번역하고 번역된 영어 웹문서를 다시 한국어로 자동번역하는 방법이다. 이러한 독한 자동번역 방법을 그림으로 그리면 다음과 같다:



【그림 1】 인터넷 독한 자동번역 방법

독한 자동번역을 직접적으로 만들 수도 있지만 이렇게 독영-영한 자동번역시스템을 연결하여 독일어를 한국어로 번역하는 데는 다음과 같은 장점이 있다:

(1) 기존의 자동번역 시스템들을 활용하여 고비용의 독한 자동번역 시스템 개발비를 절약할 수 있다.
(2) 기존에 구축되어 있는 독영, 영한의 언어학적 자원들(사전, 규칙)을 활용하여 독한 자동번역 시스템을 구축함으로써 전산시스템에 직접

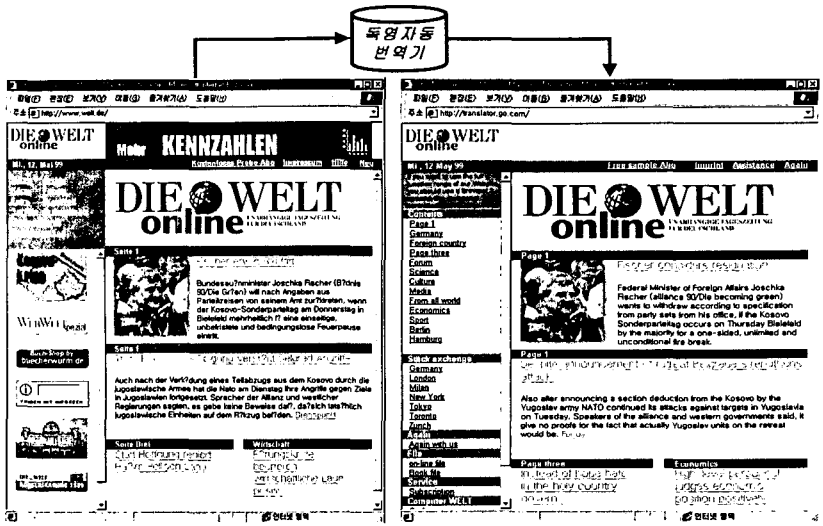
활용할 수 있는 독한 전산 사전 및 전산 문법의 중요성을 독어학 연구가들에게도 확인시킬 수 있다.

(3) 비록 한국어의 번역품질이 낮을지라도 독일어에 문외한인 일반인들에게도 대용량의 독일어 문서를 초벌 번역하여 줄 수 있다.

2.1. 독영 자동번역 방법

독일어-영어간의 자동번역에 관한 연구는 지속적인 이론적 발전이 있었을 뿐만 아니라 자동번역 제품까지도 나와 판매되고 있는 실정이다. 대부분의 독일어-영어를 대상으로 하는 자동번역 시스템들은 불어, 스페인어와 같은 유럽언어들을 포함하는 다국어 자동번역 시스템을 목표로 만들어져 왔다.

이러한 독일어-영어 언어쌍이 포함된 자동번역에 대한 연구나 제품으로는 LOGOS, SYSTRAN, Globalink 등이 있으며 번역률은 90%이상이다. 이러한 실례는 다음과 같은 다국어 자동번역 시스템인 SYSTRAN의 결과물로부터 알 수 있다.



【그림 2】 SYSTRAN에 의해 독어를 영어로 자동번역한 결과

독일어와 영어간의 자동번역 번역율이 90%이상인 이유는 독일어와 영어간에는 어순의 차이라든가 많은 다의어의 발생이 일어나지 않기 때문에 약간의 어순을 조정하며 단어 대 단어로 직접번역하는 방식으로 자동번역을 하여도 좋은 품질의 번역결과를 얻을 수 있기 때문이었다. 이러한 증거는 독일어와 영어간에 표준적인 어순의 차이가 거의 없음으로부터 알 수 있다.(Hawkins, 1983)

	독일어	영어
동사위치	SOV/ V-1, V-2	SVO/ V-1
전치사 유무	Postpositon/Prepositon	Preposition
수식어 관계	Numeral-Noun, Demonstrative-Noun, Possessive-Noun, Adjective-Noun, Genitive-Noun/Noun-Genitive, Relative Clause-Noun / Noun-Relative Clause	Numeral-Noun Demonstrative-Noun Possessive-Noun Adjective-Noun, Genitive-Noun/Noun-Genitive, Noun-Relative Clause

【도표 2】 독일어와 영어간의 어순 비교

독영 자동번역과 관련된 자동번역 시스템중에 SYSTRAN의 예를 들면 다음과 같은 특징을 지니고 있다(Mason & Rinsche, 1995) (권철중, 1999)

번역서비스 측면에서

(1) 검색시스템인 Altavista에서 1997년 12월부터 다국어 자동번역 서비스를 시작함.
(2) 서비스 언어로는 영어, 불어, 독어, 스페인어, 이태리어, 포르투갈어이다
(3) 1일에 500,000문서를 번역하여 줌(1998년 5월)
(4) 문서의 종류는 40%의 웹문서와 60%의 일반 텍스트 문서임

언어학적 측면에서

(1) 번역방식은 직접번역과 변환방식의 혼합, 완전자동번역 방식
(2) 번역대상문서의 종류는 기술서류나 매뉴얼 등

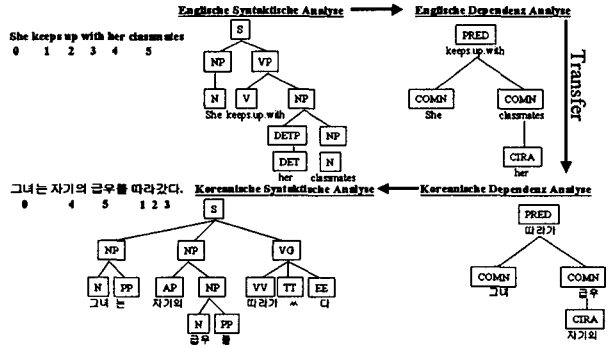
(3) 번역사전에는 어휘, 통사, 의미적 정보를 포함함
(4) 형태소, 통사 모호성 해결을 위한 100개 정도의 특수 동형이의어 처리과정을 보유함
(5) 500개의 의미범주를 가진 의미체계를 이용함
(6) 번역률은 언어별로 독영(80.0%), 러영(95.4%), 불영(86.3%), 스페인어-영어(68.9%)임.

2.2. 영한 자동번역 시스템

영어-한국어간의 자동번역에 관한 연구는 지속적인 이론적 발전뿐만 아니라 자동번역 제품까지도 나와 판매되고 있는 실정이다.

영한 자동번역에 대한 연구나 제품으로는 FromTo/EK(한국전자통신연구원), 인가이드(L&I), 앙꼬르(IBM), 트래니(언어공학연구소)등이 있으며 번역률은 40% 정도이다. 그러나 웹문서를 대상으로 한 온전한 영한 자동번역 시스템은 FromTo/Web-EK(한국전자통신연구원)밖에 없다.

한국전자통신연구원에서 추진되었던 FromTo/Web-EK에서의 전체적인 자동번역 내부과정은 다음과 같다(Choi et.al, 1999).



【그림 3】 FromTo/Web-EK의 영한 자동번역 내부과정

영한 자동번역은 영어분석, 영한변환, 한국어생성과정으로 이루어지는데

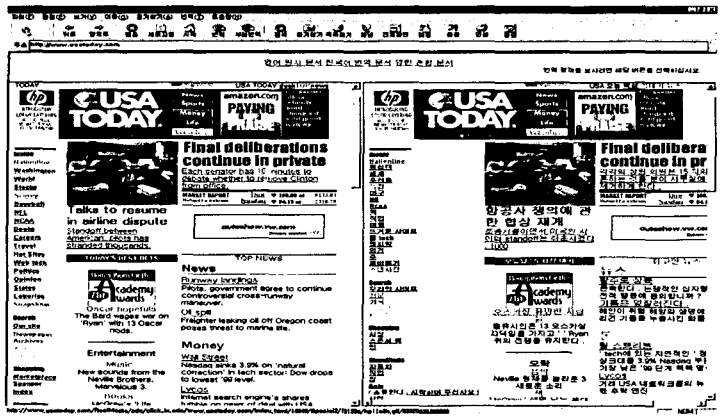
영어분석에서는 영어의 구구조를 파악하여 영어 의존구조를 만드는 일을 하며 변환에서는 영어 의존구조를 한국어 의존구조로 바꾸고 영어표현에 대응되는 한국어 대역표현을 만드는 작업을 한다. 생성에서는 한국어 의존구조를 한국어 구구조에 맞게 변형시키는 작업을 한 후 형태소 생성기에서 한국어 문장을 생성하는 작업을 하게 된다. 영한 자동번역은 전형적인 변환기반 규칙자동번역이다. 위에서 영어 단어 밑에 붙은 번호는 그 단어가 문장에서 나타나는 위치를 나타내며 이들이 한국어 대역어에서 조합 혹은 삭제 과정을 거쳐 한국어 어순에 맞는 위치에 놓이는 것을 알 수 있다. 영어와 한국어간의 어순차이는 다음과 같다. (Hawkins, 1983)

도표[3]으로부터 영한사이의 어순에 대해 알 수 있는 것은 문장성분의 어순차이가 반대로 되어 있다는 것을 알 수 있다. 즉 동사위치나 관계절 등이 그 예이다. 이러한 어순의 차이는 웹문서의 HTML Tag라든가 인용부호와 같은 기호들의 번역에도 많은 영향을 미친다.

	영어	한국어
동사위치	SVO/ V-1	rigid SOV
전치사 유무	Preposition	Postposition
수식어 관계	Numeral-Noun Demonstrative-Noun Possessive-Noun Adjective-Noun, Genitive-Noun/Noun-Genitive, Noun-Relative Clause	Numeral-Noun Demonstrative-Noun Possessive-Noun Adjective-Noun, Genitive-Noun, Relative Clause-Noun

【도표 3】 영어와 한국어간의 어순 비교

한국전자통신연구원에서 개발된 FromTo / Web-EK에 의한 영한 자동번역의 웹문서 결과물의 실례는 다음과 같다:



【그림 4】 FromTo/Web-EK에 의해 영한 자동번역 결과

3. 대조 언어학적 관점에서의 독한 자동번역의 문제와 본 논문에서의 해결방법

독일어와 한국어를 직접적으로 연결하는 독한 자동번역 연구는 언어학적인 관점에서 이론적으로 언급된 바 있었지만(최승권, 1996) (이민행 외, 1998) 실제로 작동하는 시스템 차원의 연구는 그동안 없었다. 본 절에서는 대조언어학적인 관점에서 독일어와 한국어간의 언어적인 차이를 제시하고 독한 자동번역시스템에서 해결했던 방법을 제안하고자 한다.

3.1. 범주적 차이

▷ 문제점

범주적 차이는 일대일의 조합적 번역을 위배하기 때문에 번역의 품질을 떨어뜨리는 결과를 초래한다. 이러한 예는 다음과 같은 명사구에서 발견할 수 있다.

독일어	한국어
ADJ-N(schrittweise Öffnung)	ADJ-N(단계적 개방)
ADJ-N(hohes Wachstum)	N-N(고도 성장)
DET-N(keine Videoüberwachung)	PP-FV-NEG-NSUFF(비디오감시를 하지 않음)
DET-N-DET-N(die Gerwerkschaft der Polizei)	N-N(경찰노조)

【도표 4】 독한 범주적 차이

▷ 해결방법

범주적 차이를 일으키는 단어군은 복합단위사전에 입력하여 범주적 차이도 해소하며 번역품질도 높이도록 하였다. 여기서 복합단위 사전은 숙어, 복합명사를 포함하는 사전을 의미한다. 예를 들어 'hohes Wachstum'이 영어로 'rapid growth'로 번역되면 'rapid growth'에 대한 한국어 대역은 다음과 같이 복합단위 사전에 입력이 되어 범주적 차이를 극복할 수 있었다.

rapid_growth@NOUN
[(etype comm) (sem ass) (kpos noun) (kroot 고도_성장) (kcode nn00001)]
설명: rapid_growth는 명사이며 보통명사(etype comm)이고 경제현상(sem ass)의 의미가 있고 한국어 대역품사는 명사(kpos noun)이고 한국어 대역어는 '고도 성장'(kroot 고도_성장)이고 한국어 형태소 생성을 위한 코드는 nn00001(kcode nn00001)이다

3.2. 어휘적 차이

▷ 문제점

어휘적 차이의 문제는 대역어의 올바른 어휘선택의 문제로써 번역에서 가장 해결하기 어려운 문제이며 번역품질을 떨어뜨리는 주된 원인이다. 비단 독일어와 한국어간의 대조언어학적인 문제는 아니고 자동번역의 근본적인 문제라고 할 수 있다. 이러한 예는 다음과 같은 명사구에서 발견할 수 있다.

독일어	한국어
ein besonders schwerer Verstoß	특히 심한 위반
eine schwere Arbeit	곤란한 일
ein schwerer Beutel	무거운 지갑

【도표 5】 독한 어휘적 차이

▷ 해결방법

어휘적 차이는 다의어를 해결하기 위한 어휘규칙에 의해 해결하였다. 어휘규칙은 언어간 변환과정에서 상호 연관되는 노드간의 머리어휘의 공기관계를 어휘의미와 관련하여 언급함으로써 해결할 수 있었다. 위의 예제에서 보인 'schwer'가 영어로 'heavy'로 번역되었는데 이를 영한 자동번역시스템에서 heavy에 대해 올바른 한국어로 번역하는 어휘규칙을 보이면 다음과 같다:

```
## heavy
comn-mod-tr {
  (comn! cira! ~ch)
  _with {        cira: head == [heavy];        }
_action {
    if (comn:sem == [교통]) then {
    cira:kroot : = [혼잡한]; cira:kcode : = [nn00001]; } else {
    if(comn:sem == [사회현상])_then {
    cira:kroot : = [심한]; cira:kcode : = [nn00001]; } else {
    if(comn:sem == [활동])_then {
    cira:kroot : = [곤란한]; cira:kcode : = [nn00001]; } else {
    cira:kroot : = [무거운]; cira:kcode : = [nn00001]; } } }
  }
  (comn cira ch) }
```

설명: comn-mod-tr이라는 규칙은 임의의 노드 ch를 가진 수식어 cira가 피수식어 comn를 수식하는 어휘규칙인데 cira의 머리어 head가 heavy일

때 그것의 피수식어가 '교통'의 의미를 가질 때 heavy는 '혼잡한'이란 대역어가 되고 피수식어가 '사회현상'의 의미를 가질 때 heavy는 '심한'이란 대역어가 되고 피수식어가 '활동'의 의미를 가질 때 heavy는 '곤란한'이란 대역어가 되고 기타의 의미에서는 heavy의 대역어는 '무거운'이 된다는 어휘규칙이다.

3.3. 표준 어순의 차이

▷ 문제점

독일어와 한국어 사이에는 다음과 같은 표준적인 어순의 차이가 존재한다.

	독일어	한국어
동사위치	SOV, V-1, V-2	엄격한 SOV
전치사 유무	Postpositon / Preposition	Postposition
수식어 관계	Numeral-Noun, Demonstrative-Noun, Possessive-Noun, Adjective-Noun, Genitive-Noun/Noun-Genitive, Relative Clause-Noun / Noun-Relative Clause	Numeral-Noun Demonstrative-Noun Possessive-Noun Adjective-Noun, Genitive-Noun, Relative Clause-Noun

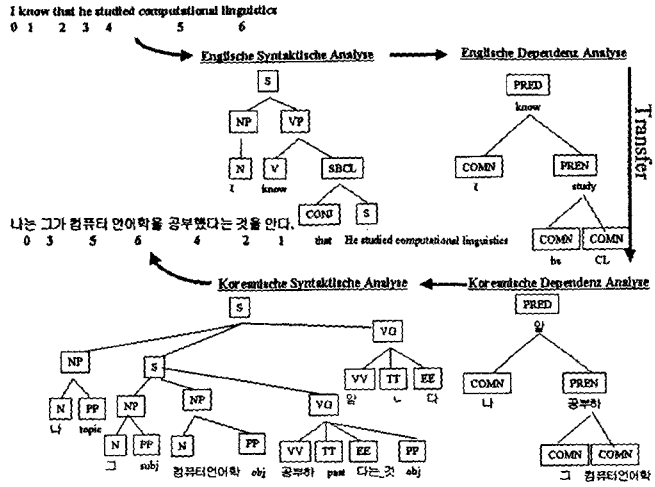
【도표 6】 독한 표준 어순의 차이

▷ 해결방법

독일어와 영어의 어순의 차이는 대부분 특이한 경우를 제외하고는 단어대 단어 번역으로 독일어가 영어로 번역되기 때문에 독영 자동번역기에서의 어순의 차이는 그리 크지 않았다. 예를 들어 다음의 독일어 문장을 독영 자동번역기가 영어로 번역한 결과를 보면 다음과 같았다:

Ich weiß, daß er Computerlinguistik studiert hat.
=> I know that he studied computational linguistics.

위의 영어 문장을 한국어로 자동번역하기 위해서는 영한 어순이 해결되어야 했다. 영어와 한국어간의 어순의 차이는 영어 구구조를 영어 의존구조로 변형시키고 영어 의존구조를 다시 한국어 구구조로 변형시키는 과정에서 해결될 수 있었다. 이러한 과정은 다음의 그림에서 잘 나타나 있다.



【그림 5】 FromTo/Web-EK에 의해 영한 자동번역 결과

4. 번역품질 개선을 위한 두 단계 전산문법

독한 자동번역기의 번역품질은 아직 낮은 상태이다. 그 이유는 독영 자동번역기의 번역률이 약 80%에 달하고 영한 자동번역기의 번역률이 약 40%에 달하기 때문이다. 근본적으로 이렇게 번역률이 낮은 이유는 모든 언어현상을 가지고 있는 웹문서의 다음과 같은 특징들 때문이다. 본 논문에서는 웹문서의 특징으로 영어 웹문서의 특징을 예로 제시하고자 한다:

(1) 빈번하게 나타나는 긴 명사구

예) Worldata's International Lists and Databases

(2) 다양한 기호를 포함하는 문장
예) UHS/Mac (Version1.2b).
(3) 도메인이나 도메인 그룹에 따른 특수 구문
예) 날씨: 요일 night 월 일
예) 장소1 about 숫자 miles northwest of 장소2.
(4) 비문법적인 영어문장
예) We're talking about years ago before anyone heard of asbestos having any questionable properties .

위에 나열된 다양한 언어현상을 가지는 웹문서를 구문분석할 때 지금까지의 전산문법이나 자동번역엔진은 다양한 번역실패를 만들었는데 그 원인을 살펴보면 다음과 같았다:

(1) 대용량의 영어 웹문장에 대응할 높은 적용력의 구문분석규칙 부재.
(2) 구조적 모호성의 미해결.
(3) 장문의 영어 웹문장 처리 미숙.
(4) 대용량의 분석.변환 사전 엔트리의 정보 부족.
(5) 비문법적인 문장의 입력시 대응 미숙.

이상의 번역실패원인은 근본적으로 구문분석문법의 높은 적용력이 부족한데서 오는 원인들이었다. 이러한 구문분석문법의 적용력을 높이기 위해 본 논문에서는 다음과 같은 두단계 구문분석을 제안하고자 한다. (장문의 영어 웹문장을 분절하는 전산문법은 본 논문에서는 지면상 제외한다.)

4.1. 제약기반 전산문법

제약기반 전산문법이란 정형화되었거나 문법적인 영어 웹문장의 올바른 분석을 담당하며 구조적 모호성을 해결하는 전산문법을 의미한다. 제약기반 전산문법을 구축하기 위하여 고려한 사항들은 다음과 같았다:

(1) 구문분석 문법 Formalismus는 어떻게 구성할 것인가?
(2) 구문분석 규칙의 적용력을 어떻게 효율적으로 높일 것인가?
(3) 구조적 모호성은 어떻게 해결할 것인가?

위의 고려사항은 각각 다음절에서 설명된다.

4.1.1. 제약기반 전산문법의 *Formalismus*

영어를 다루기 위한 제약기반 전산문법의 Formalismus로서 영어가 형상언어 konfigurationale Sprache라는 점을 감안하여 문맥자유문법을 선택하였으며 문맥자유문법에 자질을 다룰수 있는 장치, 즉 제약 Beschränkung과 통합 Unifikation를 고려하여 확장된 문맥자유문법 Erweiterte Kontextfrei Grammatik을 제약기반 전산문법의 문법으로 만들었다. 확장된 문맥자유문법의 구성형식은 다음과 같다:

구성형식	설명
제약기반전산문법인덱스	제약기반문법을 위한 표식
문맥자유규칙	문맥자유 구구조규칙
조건제약부	구구조규칙의 올바른 적용전의 제약기술
실행정보부	구구조규칙의 올바른 적용후의 정보기술

【도표 7】 제약기반 전산문법의 구성형식

확장된 문맥자유문법의 구성형식에 따라 타동사와 명사구가 묶여 타동사구를 이루는 규칙을 기술하면 다음과 같다:

```
0:    VP      ->      VERB  NP
      {       VERB:type ** [T1]; }
      {
      /* SYNTACTIC HEAD */
```

```
        VP:root := VERB:root;
        VP:type := VERB:type;
    /* SEMANTIC HEAD */
        VP:tense := VERB:tense;
        VP:modal := VERB:modal;
    /* STATISTIC INFO */
        VP:score := [1.0];
    /* STRUCTURE INFO */
        VP:lnode := VERB:lnode;
        VP:rnode := [phr];
        VP:nega := VERB:nega;
}
```

설명: 제약기반 전산문법은 인덱스 0: 로써 파악되며 동사구 VP는 동사 VERB와 명사구 NP로 이루어진다. 이 동사는 조건으로써 동사타입 typed 은 타동사라는 T1을 가져야 한다. 이 동사구 VP는 머리범주인 동사로부터 어휘 root, 동사타입 type, 시제 tense, 양상 modal, 부정정보 nega을 전달받고 확률점수 [1.0]을 가지게 된다. 이 동사구는 머리범주인 동사로부터 좌우 접속노드 정보 lnode, rnode를 전달받게 된다.

4.1.2. 제약기반 전산문법의 적용력 향상

제약기반 전산문법의 적용력을 향상시킬 수 있는 방법은 규칙들간의 충돌이 없으며 다양한 언어현상을 설명할 수 있는 대용량의 제약기반 전산규칙을 기술하는 것이다. 이를 위해 우선적으로 고려되어야 하는 것은 전산문법의 기술의 편리성을 높여야 한다는 것이다. 이를 위해 본 제약기반 전산문법은 해당 구구조 노드의 정보를 그 노드의 머리범주로부터 일관되게 복사하는 정보복사 방법을 사용하였다. 이는 기존의 머리어주도구구조문법인 HPSG(Pollard and Sag, 1994)과 같은 머리자질이동과 동일한 방법을 사용하는 것이다.

본 논문에서의 구문노드와 그것의 머리범주간의 정보이동은 다음의 도표

와 같이 나타낼 수 있다(지면상 일부분만을 기술하였다).

Non-terminal Nodes

노드명	SYNTACTICHEAD	SEMANTICHEAD
ADVP	D-HEAD, cnode, lnode, mode	D-HEAD
ADP	D-HEAD, cnode, lnode, mode, conj_lex, para, nega, nega_lex, vh	D-HEAD
AP	D-HEAD, cnode, lnode, mode, conj_lex, para, vh	D-HEAD
CMPCL	D-HEAD, cnode, lnode, mode, conj_lex	D-HEAD
CMPR	D-HEAD, cnode, lnode, mode	D-HEAD
CNCL	D-HEAD, cnode, lnode, mode	D-HEAD
DETP	D-HEAD, cnode, lnode, mode	D-HEAD
INFP	D-HEAD, cnode, lnode, mode, (missing)	D-HEAD
INGP	D-HEAD, cnode, lnode, mode, nega, nega_lex	D-HEAD
NP	D-HEAD, cnode, lnode, mode, nega, nega_lex, vh, conj_lex, cu, dist_lex, para, pp_attach	D-HEAD
PAPP	D-HEAD, cnode, lnode, mode, missing	D-HEAD
PP	D-HEAD, cnode, lnode, mode, vh	D-HEAD
QSCL	D-HEAD, cnode, lnode, mode	D-HEAD
S	D-HEAD, cnode, lnode, mode, conj_lex, para, nega, nega_lex, passth, subj	D-HEAD

Terminal Nodes

노드명	SYNTACTICHEAD		SEMANTICHEAD
ADJ_	M-HEAD	root, root_lex, a_type, form	propnt, sem_rnk
	I-HEAD	rule, cu, score, token_id, cnode, mode, lnode	
	FRAME		
ADV_	M-HEAD	root, root_lex, form, subcat	propnt, sem_rnk
	I-HEAD	rule, cu, score, token_id, cnode, mode, lnode	
	FRAME		
AUX_	M-HEAD	form	tense
	I-HEAD	root, root_lex, rule, cu, score, token_id, cnode, mode, lnode	
	FRAME		
CONJ_	M-HEAD	conj_lex	
	I-HEAD	root, root_lex, form, subcat rule, cu, score, cnode, mode, lnode	propnt
	FRAME		

【그림 6】 제약기반 전산문법의 정보이동표

위의 도표에서 각 노드명은 그것의 머리범주노드로부터 통사적 정보와 의미적 정보를 DHEAD라는 집합에 의해 복사받고 기타 해당 노드의 구문적 정보는 위의 비단말노드 Non-terminal Nodes와 같이 기술된다. 그리고 단말노드 Terminal Nodes는 상위로 보낼 정보인 MHEAD자질정보와 자체적으로 해결되어야 할 정보인 IHEAD와 상호다른 노드와의 정보 제약을 위한 FRAME으로 기술된다.

4.1.3. 제약기반 전산문법의 구조적 모호성 해결

구조적 모호성이란 동일한 문장에 대해 한 개 이상의 규칙이 적용됨으로써 한 개이상의 분석결과가 나오는 것을 말한다. 구조적 모호성 해결은 전산 언어학에서 가장 해결하기 어려우며 꼭 해결해야할 문제이다. 구조적 모호성을 해결하기 위해 여러 가지 전산학적인 해결책이 소개되기는 하였지만(Franz, 1992) 아직도 해결하여야 할 것이 많은 실정이다.

본 논문에서도 완전한 해결책은 아니지만 영어 구문분석에서의 구조적 모호성을 해결하기 위해 다양한 언어학적 정보를 이용하여 해결하였는데 이들을 열거하면 다음과 같다:

가) 좌우접속 구조정보에 의한 해결

각 구문구조에서의 구성성분이 좌우로 접속된 구조정보를 가지게 함으로써 잘못 구성되는 구문구조의 모호성을 해결할 수 있었다. 좌우로 접속되는 구조정보의 자질은 다음과 같은 도표로 나타낼 수 있다:

좌우접속 자질		좌우접속 값
좌접속 자질	lnode	nil(없음), phr(구), cla(절)
우접속 자질	rnode	nil(없음), phr(구), cla(절)

【도표 8】 제약기반 전산문법의 좌우접속 구조정보표

위의 좌우접속 자질값에 의해 다음의 동사구 규칙은 모호성이 없어지게 된다.

예) NP[0] -> NP[1] INFP {NP[0]:rnode := [cla]}
VP -> VERB NP {NP:rnode ** [nil phr cla]}
VP -> VERB NP INFP {NP:rnode ** [nil phr]}
설명) NP[0]의 우접속 자질 rnode은 값으로 절 cla을 가지는데
첫 번째 동사구에서는 명사구의 우접속 자질로 절를 허용하지만 두 번째 동사구에서는 명사구의 우접속 자질로 절을 허용하지 않기 때문에 명사구 NP[0]와 같은 부정사구를 가지는 명사구는 세 번째와 같은 동사구에서 명사구로 실현될 수 없다.

나) 동사유형정보에 의한 해결

동사유형은 동사를 분류하기 위하여 동사의 필수성분의 수와 형태를 기

반으로 만든 것으로 제약에 사용함으로써 구조적 모호성 해결에 활용할 수 있다. 동사유형은 다음의 도표와 같다:

성분 / 형식	0 (없음)	1 (명사구)	2 (부정사구)	3 (to부정사구)	4 (동명사구)	5 (절)	6 (형용사구)	7 (분사구문)	8 (부사구/전치사구)
I(SV)	I0								
L(SVC)		I1		L3	L4	L5	L6		L8
T(SVO)		T1	T2	T3	T4	T5			
D(SVOO)		D1				D5			
X(SVOC)		X1	X2	X3	X4		X6	X7	X8

【도표 9】 제약기반 전산문법의 동사유형정보 표

위의 동사유형 도표에 의해 동일한 문장구조에서 동사의 유형에 따라 구조분석이 달라 질 수 있는 것을 다음의 예문을 통해 알 수 있다:

예) I vp(vp(call you) infp(to meet him)). (call 동사타입 T1)
I vp(want you infp(to leave)). (want 동사타입 X3)

다) 사전의 전치사 필수성분 정보에 의한 해결

전치사 접속에 따른 구조적 모호성은 pp-attachment로써 널리 알려진 구조적 모호성이다. 이러한 pp-attachment를 해결하기 위해 본 연구에서는 사전에 등록된 전치사 필수성분 정보를 이용하여 pp-attachment를 해결하였다. 다음의 규칙들이 pp-attachment의 해결을 보인다:

예) NP[0] -> NP[1] PP {NP[0]:pp_attach := PP:preplex} (1)
VP -> VERB NP {VERB:arg2_prep != NP:pp_attach} (2)
VP -> VERB NP PP (3)

설명) 명사구 NP[0]의 pp_attach의 값은 전치사구 PP의 전치사 어휘 preplex로터 부여받는다(1) 동사구에서 동사가 두 번째 성분에 전치사를 가지고 있고 명사구에 전치사가 올라올 때(2) 그 두 개가 일치하면 두 번째 동사구가 아닌 세 번째 동사구가 된다는 의미임(3).

4.2. 오류허용 전산문법

오류허용 전산문법이란 비정형화되었거나 비문법적인 영어 웹문장 또는 분석실패한 영어 웹문장의 가능한 올바른 번역을 담당하는 전산문법을 의미한다. 오류허용 전산문법은 제약기반 전산문법이 적용되었지만 분석 실패한 결과물 중에서 부분적으로 성공한 구구조들을 가지고 다시 전체 구문분석나무를 재구성하는 문법이다.

4.2.1. 오류허용 전산문법의 Formalismus

오류허용 전산문법의 Formalismus는 제약기반 전산문법의 Formalismus와 동일한 형식이다. 그러나 제약기반 전산문법이 올바른 구구조 분석을 위해 제약조건을 주는 반면 오류허용 전산문법은 그러한 제약조건을 주지 않는 다는 것이 제약기반 전산문법의 구성형식과 차이나는 점이다.

오류허용 전산문법의 구성형식은 다음과 같다:

구성형식	설명
오류허용전산문법인덱스	오류허용문법을 위한 표식
문맥자유규칙	문맥자유 구구조규칙
조건제약부	없음
실행정보부	구구조규칙의 올바른 적용후의 정보기술

【도표 10】 오류허용 전산문법의 구성형식

오류허용 전산문법을 위한 확장된 문맥자유문법의 구성형식에 따라 임의

의 동사와 하나의 명사구가 묶여 동사구를 이루는 규칙을 기술하면 다음과 같다:

```
1:    VP      ->      VERB  NP
      { }
      {
      /* SYNTACTIC HEAD */
            VP:root := VERB:root;
            VP:type := VERB:type;
      /* SEMANTIC HEAD */
            VP:tense := VERB:tense;
            VP:modal := VERB:modal;
      /* STATISTIC INFO */
            VP:score := [1.0];
      /* STRUCTURE INFO */
            VP:lnode := VERB:lnode;
            VP:rnode := [phr];
            VP:nega := VERB:nega;     }
```

설명: 오류허용 전산문법은 인덱스 1: 로써 파악되며 동사구 VP는 동사 VERB와 하나의 명사구 NP로 이루어지는데 이 동사는 조건을 가지지 않음으로써 임의의 동사가 올 수 있다. 이 동사구 VP는 머리범주인 동사로부터 어휘 root, 동사타입 type, 시제 tense, 양상 modal, 부정정보 nega을 전달받고 확률점수 [1.0]을 가지게 된다. 이 동사구는 머리범주인 동사로부터 좌우 접속노드 정보 lnode, rnode를 전달받게 된다.

위에서 소개된 동사구 규칙에 따르면 타이핑 오류라든가 완전한 문장을 이루지 못한 비문법적인 다음과 같은 문장도 구문분석을 수행후 올바른 번역을 할 수 있었다.

예문) I gave a book. => 나는 책을 주었다.
He sleeps a room => 나는 방에(에서) 잔다.

첫 번째 문장은 give가 여격동사임에도 불구하고 오류허용 규칙에 의해 묶여 번역이 성공적으로 수행된 결과이며 두 번째 문장은 sleep이 자동사임에도 불구하고 타동사구문처럼 묶였다가 후에 격을 가지지 않는 명사구에 기본적으로 붙는 한국어 격조사 '에(에서)'가 붙어 성공적으로 번역된 문장이다.

5. 실험 결과

본 절에서는 독한 자동번역을 위해 사용된 영한 자동번역기의 번역지식의 크기와 독한 자동번역 결과를 제시하고자 한다. 영한 자동번역기에서 사용한 번역지식, 즉 번역규칙과 번역사전의 크기는 다음과 같았다:

규 칙			사 전	
영어구문분석규칙	제약기반 전산문법규칙	447개	영어분석사전	46,042엔트리
	오류허용 전산문법규칙	81개		
영어의존분석규칙		367개	영한 복합단위사전	19,618엔트리
영한변환규칙	어휘변환규칙	169개	영한변환사전	45,934엔트리
	구조변환규칙	86개		
한국어생성규칙		64개		

【도표 11】 번역규칙과 번역사전의 크기

독일어 원문에 대해 전문번역가의 번역과 본 논문에서 시도한 영어를 중간언어로 한 실용적인 독한자동번역 시스템에 의한 자동번역의 번역결과를 비교하면 다음과 같았다:

독일어원문	전문번역가 번역	기계번역시스템
führende Industieländer	주도적 선진국	주도하는 산업국가들
schrittweise Öffnung	단계적 개방	단계적 열림
andauemder Exportboom	지속하는 수출붐	지속하는 수출붐
hohes Wachstum	고도 성장	고도 성장
fremde Arbeitskraft	외국인 노동력	외국의 노동력/생소한 노동력
ausländische Arbeiter	외국인 노동자	외국적 노동자
starke Nachfrage	많은 수요	강한 수요
steigende Produktivität	상승하는 생산력	상승하는 생산력
eine staatlich beschlossene Sache	국가적으로 결정된 사항	국가적 결정된 물건
der vereinbarte offizielle Start	합의된 공식 출발	합의된 공식적 출발

【그림 7】 전문번역가와 독한 자동번역기의 번역결과 비교

영어를 중간언어로 한 실용적인 독한자동번역기를 가지고 99년5월12일자 "Die Welt"지의 첫면을 번역하였는데 그 결과는 다음과 같았다:

【그림 8】 99년5월12일자 "Die Welt"지의 번역결과

6. 결론

본 논문에서의 독한 자동번역 방법은 기존의 자원을 활용하는 방법이었는데 기존의 독영 자동번역 시스템과 영한 자동번역 시스템을 연결하는 방법이었다. 본 논문의 방법론에 의해 얻을 수 있는 장점과 단점을 요약하면 다음과 같다.

■ 장점

가) 자국어와 영어로 번역을 시도하고 있는 모든 인터넷 자동번역 시스템을 본 인터넷 영한 자동번역시스템과 연결시킴으로써 전혀 모르는 제3외국어의 정보도 손쉽게 얻을 수 있다.

나) 이러한 방법으로써 다국어 자동번역이 한층 손쉽게 이루어질 수 있다.

다) 본 방법론을 인터넷 한영 자동번역으로 확장하여 영어와 자국어와의 번역을 시도하고 있는 모든 자동번역 시스템을 연동시킴으로써 외국어 장벽을 한층 낮출 수 있다.

■ 단점

가) 번역의 품질이 단계적으로 낮아진다 (독일어와 한국어간의 언어학적 문화적 차이가 영어에 의해 간접적으로 한국어에 전달됨으로써 번역품질이 영어에 의해 낮아질 수 있다.)

나) 영한 자동번역의 번역품질이 낮음으로써 독일어 원문의 의미가 많이 손상되어 한국어로 파악된다.

향후 영한자동번역기의 번역엔진 및 번역지식의 개선이 이루어질 수록 본 독한 자동번역의 번역품질은 계속 개선이 될 것이다. 하지만 이러한 간접적인 방법보다는 독일어와 한국어를 직접적으로 연결하는 독한 자동번역기의 개발이 무엇보다 앞으로 연구되어야 할 부분이라고 생각한다. 이를 위하여 국내에서의 독어학 관련 학자들은 독일어와 한국어간의 언어학적 대조연구를 통한 독일어와 한국어의 다양한 언어현상의 차이를 밝히는 작

업을 한층 더 연구하여야 할 것이라고 생각된다.

참 고 문 헌

권철중(1999), "자동번역 R&D 동향", 제1회 자동번역 워크샵, 한국 전자통신연구원, 대전.

이민행/최승권/최경은(1998), "독-한 명사구 기계번역시스템의 구축", 언어와 정보, 한국언어정보학회 제2권1호, 79-105.

최승권(1996), "한국어-독일어 자동번역", 독일문학, 한국독어독문학회 37권 1호, pp.323-337.

Choi,S.K./T.W.Kim/S.H.Yuh/H.M.Jung/C.M.Sim/S.K.Park(1999), "English-to-Korean Web Translator: "FromTo/Web-EK"", In: Proceedings of MT SUMMIT VII, Singapore.

Franz,A.(1992), Ambiguity Resolution in Natural Language Processing. CMT, Carnegie Mellon University.

Hawkins,J.A.(1983), Word Order Universals. Academic Press.

Mason,J./A.Rinsche(1995), Ovum Evaluates - Translation Technology Products. Ovum Ltd.

Pollard,C./I.Sag(1994), Head-Driven Phrase Structure Grammar. Studies in Contemporary Linguistics. The University of Chicago Press, Chicago & London.

Zusammenfassung

Deutsch-Koreanische maschinelle Übersetzung auf Internet durch Kopplung von Deutsch-Englische und Englisch-Koreanische maschinelle Übersetzungssysteme

Choi, Sung-Kwon(ETRI)

Bei der maschinellen Übersetzung(=MÜ) übernimmt ein Computer die Aufgabe, Texte von einer natürlichen Sprache in eine andere automatisch zu übertragen. Die meisten MÜ-Systeme wurden bis in die 90er Jahre hauptsächlich von grossen Firmen benutzt, die viele kostenaufwendige Übersetzungsarbeiten zu bewältigen hatten. Aber die schnelle Verbreitung des Internets in den Alltag vieler Menschen hat die Situation radikal geändert, und zwar so, daß ein MÜ-System viel häufiger verwendet wird und es auch bei individuellem Bedarf Anwendung findet. Mit zunehmendem Volumen der zu übersetzenden Texte im Internet wird die Nachfrage nach einem MÜ-System immer grösser.

Ziel der vorliegenden Arbeit ist es, ein Verfahren der maschinellen Übersetzung der deutschen WWW-Texte ins Koreanische zu beschreiben.

Dazu habe ich versucht, zwei unterschiedliche MÜ-Systeme zu koppeln. Das Potential dieser Methode ist vor allem deswegen zu untersuchen, weil die sogenannte Lingware für das Koreanische und Deutsche, wie z.B. Lexikon, Grammatik, Transfer-Lexikon, nur begrenzt zur Verfügung steht.

Das erste System ist das MÜ-System SYSTRAN für das Sprachpaar Deutsch-Englisch, das die deutschen WWW-Texte ins Englische

übersetzt. Das zweite ist ein MÜ-System für das Sprachpaar Englisch-Deutsch, das die vom Deutschen ins Englische übersetzten Texte ins Koreanische übersetzt.

Bei der Evaluierung der Übersetzungen vom Deutschen ins Koreanische konnten etwa 30% der Ausgaben als korrekt bewertet werden.