다국어 질의응답을 위한 한국어 해석 시스템 설계 및 구현

강원석† · 황도삼††

요 약

다국어 질의 응답 시스템은 여러 언어의 질의에 대한 응답을 하는 시스템이다. LASSO 시스템은 다국어 질의응답 시스템 중의 하나이다. 본 논문은 LASSO 시스템을 위한 한국어 해석 시스템의 설계 및 구현에 관한 것이다. 질의 응답을 위한 한국어 해석 시스템은 한국어 질의를 처리할 수 있는 대화체 처리 기술이 필요하다. 그리고 다양한 분야의 질의에 대한 응답을 할 수 있는 범용의 시스템이어야 한다. 본 논문의 한국어 해석 시스템은 이와 같은 사항을 만족하기 위하여 심도 깊은 대화체 처리 기술보다 실용성이 높은 휴리스틱 규칙을 활용하였다. 이 시스템은 다국어 질의 응답 시스템의 한국어 인터페이스 역할을 하는 것으로 질의 응답 시스템의 목적에 맞게 설계, 구현되었다. 본 해석 시스템에 적용된 기술은 정보검색 분야와 한국어 해석 분야에 응용할 수 있다.

키워드 : 구문해석, 의미해석, 질의응답

Design and Implementation of a Korean Analysis System for Multi-lingual Query Answering

Won-Seog Kang† · Do-Sam Hwang††

ABSTRACT

Multi-lingual query answering system is the system which answers on the queries with several languages. LASSO[1] is the system that aims to answer the multi-lingual query. In this paper, we design and implement a Korean analysis system for LASSO. The Korean analysis system for query answering needs processing techniques of dialogue style. And the system must be practical and general so as to use on various domains. This system uses not dialogue processing techniques with high cost and low utility but heuristic rules with low cost and high utility. It is designed and implemented as a Korean interface of multi-lingual query answering system. The techniques of this system highly contribute to information retrieval and Korean analysis researches.

Keywords : Syntactic analysis, Semantic analysis, Query answering

1. 서 론

정보 사회에서 정보 검색은 필수적이다. 이와 같은 필요성에 따라 정보 검색에 대한 많은 연구[1,2]가 이루어지고 있다. 인터넷의 보급은 정보 검색의 필요성을 더 가중시키고 있다. 그러나 인터넷에 널려 있는 수많은 정보들은 여러 나라의 언어로 기술되어 있어 한 언어에 대한 정보 검색 시스템은 그 가치가 절하된다. LASSO[1]는 이와 같은 문제점을 해결하는 다국어 질의 응답 시스템 중 하나이다. 이 시스템은 질의에 대한 응답으로 필요로 하는 정보의 전체 텍스트를 찾아주는 대신 텍스트의 일부 문맥을 찾는 방법에 대한 연구이다.

† 정 회 원: 안동대학교 컴퓨터교육과 교수(교신저자)
†† 정 회 원: 영남대학교 컴퓨터공학전공 교수
논문접수: 2004년 6월 23일, 심사완료: 2004년 7월 13일
* 이 논문은 2001년 안동대학교의 학술연구조성비에 의하여 지원되었음

LASSO에서는 정보 검색 기법이나 정보 획득 기법이 주어진 질의에 대한 일부 문맥을 찾아주는 방법으로 그대로 적용할 수 없다고 판단하고 구문 해석에 관한 정보 검색 기법과, 심층 레벨의 자연어 처리 기법 적용 없이 질의에 대한 답을 찾는 휴리스틱 방법의 정보 획득 기법을 적용하였다. 이 방법으로 심층 레벨의 자연어 처리 기법을 적용하지 않고서도 질의 응답 시스템의 질을 향상시킬 수 있음을 보였고 다국어 질의 응답 시스템으로 시스템을 확장하기 위하여 각 언어에 대한 인터페이스를 추가하고자 하였다. 본 연구는 이 다국어 질의 응답 시스템의 한국어 인터페이스를 위하여 한영 번역에서의 한국어 해석 시스템의 설계와 구현을 시도한다.

한국어 해석 분야에서는 많은 연구가 진행되어 왔다. [3,4,5,6,7,8]등은 규칙 중심의 구문 분석을 시도하여 구문 분석의 효율 향상이나 모호성 해소를 꾀하였다. [9]의 연구는 규칙 중심의 해석을 보완하기 위하여 확률 모델의 구문 분석을 제안하여 중의성 해소를 시도하였다. [10]은 대량의 코퍼스를 활용하여 일반화된 구문 분석을 시도하였다. [11,12]는 점진적인 시스템을 목표로 학습 방법을 이용한 구문분석을 제안하였다. [13,14]는 사전정보와 함께 시소러스 정보나 코퍼스 정보를 이용하여 통합적인 격의미 해석을 시도하였다. 이와 같은 방법들은 규칙을 활용하거나, 사전이나 시소러스의 활용, 말뭉치나 통계의 이용, 또는 학습 방법을 이용하고 있다. 이 방법들은 본 연구의 배경인 다국어 질의응답 시스템에 적용하기에는 어려움이 많다. 다국어 질의 응답 시스템은 분야가 넓고 일관성이 없는 정보를 다루어야 하기 때문에 제한된 분야의 문서를 처리하는 언어 처리 기법은 효과가 높지 않다. 따라서 본 연구는 구문 해석과 격의미 해석에 심층 레벨의 자연어 처리 기법 적용 없는 휴리스틱 규칙과 의미정보를 추출하는 시소러스 도구를 활용하여 분야가 다양하고 표현의 일관성이 없는 한국어 문서의 처리를 시도하였다.

본 논문의 한국어 해석 시스템의 설계 원칙은 다음과 같다..

1) 엄격하지 않는 번역

2) 다국어질의응답을 위한 질의유형 분석

첫째로 본 한국어 해석 시스템은 엄격하지 않는 번역을 원칙으로 적용한다. 엄격하지 않는 번역(loose translation)은 다국어 질의응답 시스템의 질의/응답에 차질을 주지 않는 한도의 번역을 허용하는 것을 말한다. 즉, 한국어 해석의 질 향상이 아닌 질의 응답 시스템의 성능을 염두에 둔 해석에 초점을 두고 있다는 것을 의미한다.

둘째로 본 한국어 해석 시스템은 최종 목표인 다국어 질의 응답시스템 LASSO의 질의 응답을 위해 질의의 유형 분석을 한다. 질의 유형 분석은 입력이 되는 한국어 질의 유형 기준을 따르지 않고 LASSO 유형 기준을 따라 분류하고 질의를 해석한다. 이는 한국어와 영어가 질의를 표현하는 방법에 있어 많은 차이가 있기 때문이다[17]. 한국어 유형기준을 따라 분석한다면 한영 번역 과정을 통해 문제를 야기할지도 모른다. 본 논문에서는 그와 같은 문제를 줄이기 위하여 한국어 해석시에 영어 질의의 유형 기준을 잣대로 삼아 한국어 질의가 어느 영어 질의 유형에 해당하는지를 밝힌다.

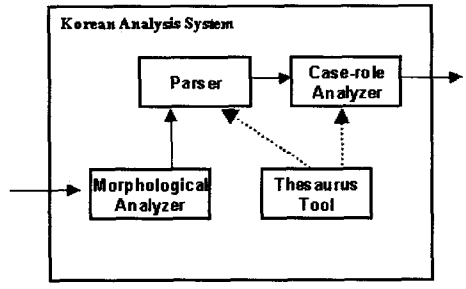

그림 1. 한국어 해석 시스템의 구조도

본 논문의 한국어 해석 시스템의 구조는 그림 1과 같다. 이 시스템은 LASSO를 위한 한영번역의 해석 시스템으로 세 단계의 해석으로, 형태소 해석, 구문 해석, 격 해석으로 구성된다. 형태소 해석을 위해 [15]에서 개발된 형태소 해석기를 사용한다. 파서는 앞에서 설명한 원칙을 적용하여 파스 트리를 구성하고 질의유형을 파악한다. 격 해석기는 영어 전치사 생성을 할 수 있도록 격해석을 하여 번역의 다음 단계로 정보를 넘긴다. 한국어 해석 시스템은 세 단계의 시스템과 함께 단어의 의미속성 정보를 제공하는 시소러스 도구가 포함된다. 이 도구는 파서와 격 해석기에 단어의 정보를 제공한다. 이 도구는 [16]의 도구를 확장한 것이다.

2. 구문 해석

한국어는 첨가어이기 때문에 구성요소의 구조적

인 순서와 위치에 민감하지 않다[3,11,17]. 따라서 한국어 구문해석은 단어나 구의 구조적인 분석보다 기능적인 분석에 집중해야 한다. 즉, 단어나 구의 기능을 나타내는 조사나 어미의 분석에 초점을 맞추어야 한다. 그렇지만 조사나 어미의 세부분석은 의미 레벨의 격해석 단계까지 언급하게 된다. 격해석에 대한 것은 시스템의 다음 단계에서 다루게 되므로 이 단계에서는 일반적이고 구문적인 기능을 표현하는 파스 트리를 만드는 것에 목표를 둔다.

구문 해석기는 다음과 같은 네 단계의 처리를 거친다.

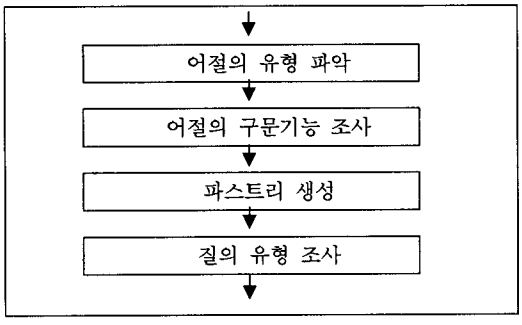

그림 2. 구문해석기

첫째 단계는 어절의 유형을 파악하는 것이다. 유형 결정은 형태소 해석의 결과를 토대로 유형을 결정한다. 즉, 헤드의 품사와 첨가어의 종류에 따라 어절 유형이 결정된다. 표 1은 각 어절의 유형을 기술하였다. 어절의 헤드는 어절의 주 의미를 담고 있는 명사나 동사를 말하고 첨가어는 헤드에 의존되는 조사나 어미 등을 말한다.

표 1. 어절의 유형

유형	헤드	첨가어	예
HNOUN	명사류	(조사들)	동물+은
HNXS	명사류	동사파생접미사+어미	공부+하+다
HVERB	동사류	어미	달리+다
HBEVERB	명사류	서술격조사+어미	동물+이+다
HPTN	동사류	명사형어미+조사	살+ㅁ+은
HMA	부사		너무
HMM	관형사		그

두 번째 단계는 어절의 구문적인 기능을 조사한다. 표 2는 어절의 가능한 구문격을 기술하고 있다. 구문해석에 정의된 구문격은 다국어 질의응답 시스템의 처리를 고려하여 설계되었다. 구문해석기는 어절의 구문격을 파악하기 위하여 어절의 헤드, 조사나 어미, 피수식어의 세 요소의 정보를 검토한다. 피수식어는 어절이 수식하는 대상으로 서술어이거나 다른 체언 등이 될 수 있다. COMP 격의 경우 피수식어는 "이다"나 "되다"와 같은 서술어가 될 것이다. 관형격의 경우 어절은 전치사구 "of"로 번역될 것이고 결합격의 경우는 "and"에 해당된다. 서술격의 경우는 영어의 "be" 동사와 유사하다.

표 2. 구문격

격이름	조사나어미	예
주격(SUBJ)	가,이,에서,서	동물+이
목적격(OBJ)	을,를	나무+를
보격(COMP)	이,가	성인+이
관형격(ADNM)	의	나무+의
호격(VOC)	야,아,여,이야,이여	철수+야
결합격(JUNC)	와,과,며,랑	철수+와
서술격(PRED)	이다	동물+이+다

세 번째 단계는 파스 트리를 구성하는 단계이다. 구문 해석기는 어절을 트리로 만들어 간다. 파싱에 적용되는 일반적인 원칙은 다음과 같다.

1) no-crossing principle
2) locality principle

구문해석기는 대부분 1과 2의 원칙을 준수한다. 본 구문해석기도 이 원칙을 기초하여 그림 3과 같은 휴리스틱 규칙을 적용한다.

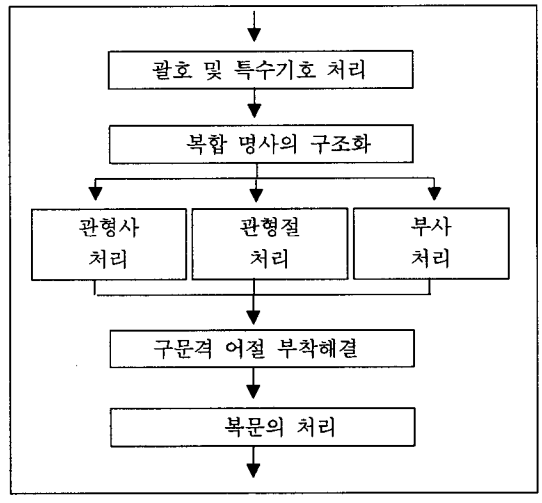

그림 3. 파스 트리 생성

파스 트리 생성의 첫 단계는 괄호와 특수기호 처리이다. 괄호의 내용이 하나의 문장인 경우는 이 내용을 해석하여 파스트리를 구성한다. 만약 그렇지 않으면 괄호의 내용을 하나로 묶어 하나의 요소

로 둔다. 두 번째 단계는 복합명사를 만드는 과정이다. 이 단계에서는 연속적으로 나오는 명사들을 하나로 묶는다. 마지막 명사는 조사나 관계사를 수반하는 경우이다. 본 시스템의 복합명사 구성은 한국어 해석기의 설계 원칙에 의거하여 기초적인 것만 다루었다. 다음으로 관형사나 관형절, 부사의 경우로 나뉘어 처리하게 된다. 관형사나 관형절은 한국어의 특성과 locality 원칙을 고려하여 가장 가까이 따르는 명사에 붙였다. 부사는 동사, 파생동사 또는 서술격 동사에 부착한다. 여기에도 locality 원칙을 반영하였다.

다음으로 구문격의 어절 부착해결 단계이다. 어절이 어디에 부착되는지 알아보기 위해 구문해석기는 세 개의 구성요소를 검사하였다. 그 구성 요소는 어절의 헤드 정보, 조사나 어미 정보, 피수식어의 정보이다. 각 정보의 검사는 단어 어휘 식별과, 시소러스 도구를 통해 추출된 의미정보 파악으로 이루어진다. 이 단계는 격해석 단계에서 사용하는 정보를 이용하게 된다. 이 단계에서도 locality 원칙을 기초하여 표 4에 도시된 격의미 정보에 해당하는 것 가운데 가장 가까운 것을 선택하였다. 마지막으로 복문의 처리에서는 구성된 단문의 파스트리를 대상으로 문장의 관계를 검사한다. 단문의 서술어의 어미가 부사형이면 이 단문은 후속되는 동사나 파생동사, 또는 서술격조사의 동사와 연결한다. 만약 단문의 서술어의 어미가 관형형이면 후속되는 명사나 서술격조사의 어절을 찾아 연결한다.

구문 해석의 네 번째 단계는 한국어 질의의 질의 유형을 분석하는 단계이다. 질의 유형의 기준은 LASSO 시스템의 질의 유형을 따른다. 앞에서 언급한 바와 같이 한국어와 영어의 질의 표현상에 차이가 많이 있다. 이에 따른 문제를 줄이고자 질의 유형 분석의 기준을 한국어의 기준을 따르지 않고 LASSO의 기준을 따랐다. 표 3은 질의 유형을 분석하기 위한 휴리스틱 규칙을 기술하고 있다.

표3의 첫 번째 열은 한국어 질의 단어를 나타내고 두 번째 열은 그 단어의 분류태그를 표시한다. 세 번째 열은 질의의 유형을 결정하는 휴리스틱 규칙의 실마리를 나타내고 네 번째 열은 해당하는 LASSO의 질의유형이다. 질의 유형 결정 휴리스틱 함수는 다음과 같다.

find_qtype(Type$, Tag$, Clue$())

- Type$: 의문사 유형
- Tag$: 의문사의 품사 유형
- Clue$() : 각 의문사의 제한조건 검사함수 유형

표 3의 각 행에 대해 해당하는 휴리스틱 함수를 차례로 수행하여 질의 유형을 결정한다. Clue$()의 경우 각 함수의 정의는 다음과 같다.

- qc_pt(Postpt) : 의문사의 조사가 Postpt인지 검사하는 휴리스틱 함수
- qc_fn(SemFea) : 후속 명사의 속성이 SemFea인지 검사하는 휴리스틱 함수
- qc_sb(SemFea) : 주어 속성이 SemFea인지 검사하는 휴리스틱 함수
- qc_me(SemFea) : 피수식어의 속성이 SemFea인지 검사하는 휴리스틱 함수
- qc_mp(Pos) : 피수식어의 품사가 Pos인지 검사하는 휴리스틱 함수
- qc_pp(Tag) : 의문사의 부착어의 품사태그가 Tag인지 검사하는 함수

"왜"나 "언제"의 경우 Clue$()함수는 정의되지 않고 단지 Type$의 존재여부와 해당하는 Tag$의 매칭 여부만 검사하고 "무엇"의 경우는 의문사의 조사가 무엇인지를 검사하는 qc_pt() 함수를 통해 질의 유형이 결정된다.

표 3. 질의 유형 결정 휴리스틱 규칙

한국어 Type	분류태그 Tag$	실마리 Clue	질의유형
왜	지시부사		why
언제	지시부사		basic-what
무엇	지시대명사	qc_pt(의)	whose
		qc_pt(이)	basic-what
		qc_pt(으로)	what-targ
어디	지시대명사/지시부사		where, which-where
어떻게	지시부사		basic-how
몇	비서술성명사/수관형사	qc_fn(시간)	which-when
		qc_fn(수)	how-many
얼마	비서술성명사/지시대명사	qc_sb()	how-many,how-much,how-long,how-far,how-tall,how-rich,how-large
얼마나	지시부사	qc_sb() & qc_me($)	how-many,how-much,how-long,how-far,how-tall,how-rich,how-large
어떠하	지시형용사	qc_mp(N)	which
		!qc_mp(N)	basic-how
누구	인칭대명사	qc_pp(jp)|qc_pp(jcs)	who
		qc_pp(jco)|qc_pp(jca)	whom
무슨/어느/어떤	지시관형사	qc_sb()	which-who,which-where,which-when,which-what

예 문장에 대한 구문해석의 결과를 그림 4에 도시하였다. 이 문장의 파스 트리의 루트는 노드 8이다. 그리고 노드 6과 노드 2는 각각 두 개의 후손 노드를 가지고 있음을 보인다.

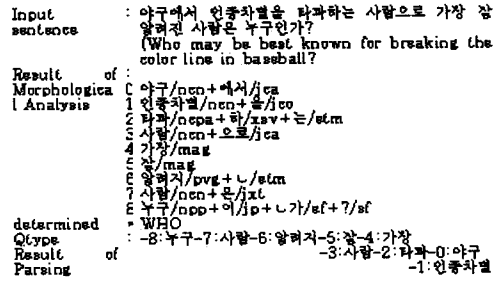

그림 4. 구문해석의 결과

3. 격의미 해석

구문 해석의 다음 단계로 격의미 해석을 하여 어절의 세부 격의미를 결정하게 된다. 세부 격의미는 목적언어인 영어의 전치사 선택의 정보가 된다. 주격어절과 목적격 어절의 경우는 전치사가 필요없기 때문에 세부 격의미를 결정하지 않아도 되나 다른 어절의 경우는 격의미 해석을 해야 한다. 격의미 해석은 표 4와 같은 격의미 결정 휴리스틱을 따른다. 구문해석 단계에서 간단히 언급했듯이 격의미를 결정하기 위하여 어절의 헤드 정보와 조사 정보, 그리고 피수식어의 의미 정보의 세 구성요소의 조건과 기타 사항을 검사해야 한다. 검사 휴리스틱 함수는 네 가지 요소의 검사로 이루어진다.

find_semrole(Postpt$, Head$, Modifiee$, Chkinf$())

- Postpt$: 격조사 유형
- Head$: 헤드의 의미속성집합 유형
- Modifiee$: 피수식어의 의미속성집합 유형
- Chkinf$() : 기타 제한조건 검사 휴리스틱함수 유형

Chkinf$() 휴리스틱 함수는 어떤 제한조건을 만족하는지를 검사하는 함수이다. 각 종류는 다음과 같다.

- rc_hd(Str) : 헤드에 Str이 포함되어 있는지 검사하는 휴리스틱 함수
- rc_me(Str) : 피수식어 정보가 Str인지 검사하는 휴리스틱 함수
- rc_pv() : 동사가 피동형인지 검사하는 휴리스틱 함수
- rc_mt(Str) : 피수식어의 영어 대역어가 Str인지 검사하는 휴리스틱 함수
- rc_pf(Str) : 문장에서 격조사의 뒤에 Str이 들어있는지 검사하는 휴리스틱 함수
- rc_fp(Str) : 문장에서 격조사 앞에 Str이 들어있는지 검사하는 휴리스틱 함수

표 4. 격의미 집합 및 결정규칙

격조사 Postpt$	헤드 Head$	피수식어 Modifiee$	빈도수	격(전치사)	기타 Chkinf$()
에	location,structure	positioning	101	loc(in)	
	location,structure	positioning	2	loc(on)	rc_mt('put')
	time		31	tim(in)	
	time-point		2	tim(at)	
	time		1	tim(after)	rc_hd('후') | rc_hd('뒤')
	organization,state,situation	rc_me('선출되')		stat(as)	
	duration		3	peri(during	rc_hd('동안')
	animate-thing	rc_me('의하')	11	subj(by)	
			48	targ(on,to,for,with,in	rc_mt($)
로/으로	organization,state,situation		57	stat(as,for,in,by)	rc_mt($)
	tool		4	inst(with)	
	measurement		12	meas(in)	
			4	targ(over,to,in)	rc_mt($)
	rc_hd('적')		4	adverb	
에게		rc_pv()	1	subj(by)	
			5	targ(for,bl)	rc_mt($)
부터,으로부터,로부터	location		2	sr-p(from)	
	time			sr-t(from)	
			1	sr(from)	
서,에서	location		58	loc(in,on)	rc_mt($)
	location,rc_hd('아래')		1	loc(under)	
	situation		51	situ(in)	
	location,structure		5	sr-p(from	rc_pf('까지')
			11	sr(from)	
와,과	animate-thing	competition	1	cmpe(with	
	animate-thing	agreement		accm(with	
까지	time		1	dstt(to)	rc_fp('에서') | rc_fp('부터')
	location		5	dst(to)	rc_fp('에서') | rc_fp('부터')
마다			1	adverb	

표 4의 첫 번째, 두 번째, 세 번째 열의 항목은 격의미에 대해 세 개의 구성요소의 제한조건을 표현하고 있다. 네 번째 열은 예 문장에서의 빈도수를 나타낸다. 예 문장은 236 문장이다. 다섯 번째

열은 그 제한조건을 만족하는 격의미와 생성될 전치사를 나타낸다. 기타 란은 각 격의미에 대한 부가 제한조건을 나타내는 것으로 Chkinf$()에 해당한다.

격의미 결정 규칙에서는 제일 먼저 격조사 "에"에 대해 검사를 한다. 한국어의 "에" 조사는 그 쓰임새가 다양하다. 장소로 사용할 수 있는가 하면 시간, 상태, 기간, 주격, 대상격 등으로 사용이 가능하다. 따라서 세부 격의미의 결정은 쉽지 않다. 만약 어절의 헤드가 장소나 구조의 의미를 가지고 있고 어절의 피수식어가 위치지정의 의미를 가지고 있으면 시스템은 전치사 "in"을 생성하는 장소 격의미로 결정한다. 만약 생성할 영어의 동사가 "put"이라면 시스템은 전치가 "on"을 생성하는 장소 격의미로 결정한다. 이와 같은 경우는 그 결정이 영어 단어가 결정될 때까지 미루어진다. 어절의 헤드가 시간의 정보를 가지고 있다면 "에"는 시간 격으로 결정된다. 만약 단어 "후"와 "뒤"가 헤드의 어절 속에 포함된다면 그 격의미는 전치사 "after"를 생성하는 시간 격의미로 세분된다. 이 휴리스틱 규칙의 적용은 표의 순서로 진행되었다. 즉, "에" 격조사에 대해서 상태, 기간, 주격, 목적격의 순서대로 격을 검사한다. 목적격의 경우 생성될 전치사가 여러 개가 가능하다. 그림 5에서 노드 2의 격의미는 "targ_to"로 결정되어 있지만 영어를 생성할 때에 영어 단어가 spend로 결정된다면 이와 결합할 수 있는 전치사 on이 생성될 것이다.

```
Input sentence:  Mercury가 1993년에 광고하는데에 얼마를 사용
                 했나?
                 (How much did Mercury spend on
                 advertising in 1993?)
Result       :   -4:사용(finen)-2:광고(targ_to)-0:Mercury(s
                 ubj)
                                        -1:1년(tim-in
                              -3:얼마(obj)
```

그림 5. 격의미 해석의 결과

조사 "로"의 경우 격의미는 헤드 정보를 이용한다. 만약 헤드가 "적"이라는 파생접미사를 가지고 있다면 그 어절은 영어의 부사로 생성될 것이다. 만약 해당하는 제한 조건이 없다면 그 어절은 디폴트 격의미 target이나 빈도수가 높은 격의미로 결정될 것이다.

일반적으로 해석시스템의 성능은 해석 방법에 달려있다. 좋은 해석 방법이 개발된다고 할지라도 의미를 구분할 수 있는 의미속성을 획득하는 도구의 지원 없이는 좋은 성능을 얻을 수가 없다. 본 논문에서는 한국어 해석에 필요한 의미속성을 획득할 수 있는 시소러스 도구를 사용한다. 이 도구는 문서에서 빈번히 출현하는 단어 8631단어에 대해 의미를 획득할 수 있다. 이 도구는 일관성이 없는 문서의 정보 처리에 중요한 의미를 부여한다. 앞으로 더 많은 단어에 대해 의미를 추출할 수 있는 도구로 확장할 예정이다.

4. 실험 및 결과

본 논문에서 우리는 LASSO에 사용된 예제 문장 236개에 대해 구문해석기와 격의미 해석기를 실험하였다. 표 5는 그 실험 결과를 나타내고 있다.

표 5. 시스템의 실험 결과

시스템	성공률	허용률	실패율
구문해석기	190/236(80%)	212/236(90%)	24/236(10%)
의미해석기	607/700(87%)	667/700(95%)	33/700(5%)
질의유형분석	175/236(74%)	219/236(93%)	17/236(7%)

표에 나타난 의미해석기의 허용률은 성공률에 세부 격의미 연장의 경우를 포함한 것이다. 예를 든다면 시스템은 "TARGET-ON" 대신에 "TARGET-IN"으로 결정한 경우이다. 본 시스템은 생성할 영어 동사가 무엇인지에 따라 세부적인 격의미가 결정되는 경우 그때까지 세부 격의미 결정을 연기한다고 하였다. 허용률은 이 경우의 실패율을 포함한 것이다.

구문해석기의 허용률은 파스트리를 구성할 때 전체 주요 구조에 영향을 주지 않는 경우를 포함한 것이다. 질의 유형 분석의 허용률은 해석결과의 유형과 실제 질의의 유형과 유사하여 그대로 사용할 수 있는 경우를 포함한 것이다. 실험 결과로 보아 본 시스템의 허용률은 상당히 고무적임을 알 수 있다.

본 논문에서 우리는 시스템의 분석과 개선을 위하여 시스템이 실패한 경우를 고찰하였다. 실패한 경우의 상당 부분이 생략, 병렬표현, 의문사 단어와 격조사 결합에 의한 의미속성 획득의 불가에 의해 일어났다. 각 경우에 대해 설명한다. 먼저 구문해석기의 경우 실패한 원인은 다음과 같았다.

- 주 동사의 생략 : 8%
- locality 원칙의 오류 : 25%
- 복합동사 미처리 : 29%
- 형태소 해석기의 오류 : 17%

이를 개선하기 위하여 생략 현상을 처리하는 과

정과 복합동사의 처리 단계가 필요하다. 이 문제는 개선의 여지가 있는 부분이다. 그러나 locality 원칙의 문제는 좀 더 깊은 단계의 처리가 필요하나 규칙성이 없는 문서의 경우에 다른 해결방법이 필요하고 앞으로의 과제이다.

격의미 해석기에서는 실패한 경우의 30%가 격조사 "에서"에서 발생하였다. 장소격으로 사용되어진 것이 시스템은 원시격으로 잘못 해석하였다. 그리고 30%는 조사 "에"에서 발생하였다. 장소격이나 시간격으로 사용된 것이 시스템은 대상격으로 구분하였다. 이 실패율을 줄이기 위해 격조사에 대한 언어적 특징을 조사하고 반영하여야 한다.

질의유형 분석의 경우 실패한 경우의 24%가 의문사 단어가 생략되어 발생하였다. 그리고 59%는 의문사의 대상 단어의 속성을 알 수 없어서 발생하였다. 이 경우 생략 처리 단계의 추가와 의문사 대상 단어의 속성 파악 프로시저를 추가하여 시스템을 개선할 수 있다.

5. 결론 및 향후과제

다양한 문서에 대한 질의 응답을 할 수 있는 다국어 질의응답 시스템의 필요가 증가하고 있다. 본 논문은 이러한 필요성에 따라 다국어 질의 응답 시스템을 위한 한국어 해석시스템을 설계하고 구현하였다. 규칙성이 없고 범위가 다양한 문서의 처리는 기존의 자연어 처리 방법을 그대로 적용할 수 없다. 본 논문에서는 이를 처리할 수 있는 휴리스틱 함수를 활용한 한국어 해석 시스템을 제안하였다. 본 시스템을 실험한 결과 구문 해석기의 경우 허용률이 90%, 의미해석기의 경우 허용률이 95%, 질의유형 분석이 93%의 결과를 얻었다. 이 연구결과는 앞으로 생략현상의 처리, 복합 동사의 처리의 연구와 격의미 결정 휴리스틱의 개선이 필요함을 보인다.

참 고 문 헌

[1] D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Girju and V. Rus, "LASSO: A Tool for Surfing the Answer Net," Proceedings of the Text Retrieval Conference (TREC-8), November, 1999.

[2] G.Salton, Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley, 1989.

[3] 서광준, 어절간의 의존관계를 이용한 한국어 파서, KAIST 석사논문, 1993.

[4] 이공주, 김재훈, 김길창, "제한된 형태의 구구조 문법에 기반한 한국어 구문분석," 정보과학회 논문지(B) 25권 4호, 1998.4.

[5] 이공주, 김재훈, "규칙에 기반한 한국어 부분 구문분석기의 구현," 정보처리학회논문지 B 제10-B권 4호, 2003.8.

[6] 이재원, 김영택, "모호성 패턴을 이용한 구문 규칙의 개선과 구문 모호성의 해소," 정보과학회 논문지(B), 25권 5호, 1998.5.

[7] 김미영, 강신재, 이종혁, "단위(Chunks)분석과 의존문법에 기반한 한국어 구문분석," 한국정보과학회 2000년 춘계학술대회 논문집, 27권 1호, 2000.

[8] 이현영, 황이규, 이용석, "문형과 단문 분할을 이용한 한국어 구문 모호성 해결," 2000년 한글및한국어 정보처리학술대회 논문집, 2000.

[9] 이공주, 김재훈, "중심어간의 공기정보를 이용한 한국어 확률 구문분석 모델," 한국정보처리학회 논문지B 제9-B권 6호, 2002.12.

[10] 윤준태, 김선호, 송만석, "전역적 연관표를 이용한 한국어 구문분석," 정보과학회논문지(B) 24권 11호, 1997.11.

[11] 최영림, 신경망을 이용한 한국어 격해석기의 구현, KAIST 석사논문, 1995.

[12] 박소영, 곽용재, 정후중, 황영숙, 임해창, "한국어 구문분석의 효율성을 개선하기 위한 구문제약규칙의 학습," 정보과학회논문지: 소프트웨어 및 응용 29권 10호, 2002.10.

[13] 양재형, 심광섭, "시소러스와 하위범주화 사전을 이용한 격 모호성 해결," 정보과학회논문지(B) 26권 9호, 1999.9

[14] 강신재, 박정혜, "대규모 말뭉치와 전산 언어사전을 이용한 의미역 결정 규칙의 구축," 정보처리학회논문지 B 제 10-B권 2호, 2003.4.

[15] 김재훈외 5인, "새 환경에 적응가능한 한국어 품사 태깅 시스템 KTAG99," 11회 한글 및 한국어정보처리 학술대회논문집, 1999.

[16] W. S. Kang and H. K. Kang, "An Effective Concept-based Text Categorization System Using the Thesaurus Tool," Proc. of 18th Int. Conf. of Computer Processing of Oriental Languages, March, 1999.

[17] W. S. Kang, Semantic Roles for English-to-Korean Machine Translation, CSMEMO-93-01, KAIST, 1993.

[18] D. Jurafsky and J. H. Martin, Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice-Hall, 2000.

강 원 석

1985 경북대학교 전자공학과 (공학학사)
1988 한국과학기술원 전산학과 (공학석사)
1995 한국과학기술원 전산학과 (공학박사)
1995~현재 안동대학교 컴퓨터교육과 교수
관심분야: 컴퓨터교육, 자연어처리, 정보검색
E-Mail: wskang@andong.ac.kr

황 도 삼

1980 홍익대학교 전자계산전공 (이학학사)
1983 연세대학교 전자계산전공 (이학석사)
1995 교토대학교 전자정보통신 (공학박사)
1996~현재 영남대학교 컴퓨터공학전공 교수
관심분야: 자연어처리, 정보검색, 컴퓨터교육
E-Mail: dshwang@yu.ac.kr