국립국어원 2008-01-21

발 간 등 록 번 호
11-1371028-000024-01

국어정보화 2단계 사업 계획 수립

국립국어원

국립국어원 2008-01-21

발 간 등 록 번 호
11-1371028-000024-01

국어정보화 2단계 사업 계획 수립

연구책임자 : 서영훈(충북대학교)

국립국어원 • 한국정보과학회

제 출 문

국립국어원장 귀하

"**국어정보화 2단계 사업 계획 수립**"에 관하여 귀 원와 체결한 연구 용역 계약에 의하여 연구최종보고서를 작성하여 제출합니다.

2008년 12월 20일

사단법인 한국정보과학회

연구책임자 : 서영훈(충북대학교)
공동연구원 : 강승식(국민대학교)
김경선(다이퀘스트)
윤애선(부산대학교)
최재웅(고려대학교)
최호섭(KISTI)

자문 위원 : 강현규(건국대학교)
심철민(파란닷컴)
박동인(KISTI)
옥철영(울산대학교)
이성헌(서울대학교)
최기선(KAIST)
홍문표(성균관대학교)

요 약 문

1. 사업명

"**국어정보화 2단계 사업 계획 수립**"

2. 연구의 목적

본 사업은 지난 10년간 수행된 21세기 세종계획을 비롯한 기존의 국어정보화 사업을 통해 구축된 자원의 응용적·활용적 관점에서 분석하고, 그동안 변화된 국어 정보 환경을 조사·분석하여, 이를 바탕으로 2010년부터 2014년까지 5년간 국가사업으로 추진할 **국어정보화의 중장기 사업 계획**을 수립하는 것을 목적으로 한다.

3. 연구의 필요성

○ **기술 환경 및 사회적**으로 시장 환경의 변화에 따라 포털 사업자들의 국어자원 구축과 응용 기술 개발의 필요성이 점차 더 높아지고 있으며, 검색 서비스의 확장을 위해서는 **신조어, 전문용어 등의 언어 자원을 지속적으로 확충**해야 하고 검색 서비스의 품질 개선을 위해 공신력 있는 **평가 세트 및 테스트 환경**을 구축할 필요가 있다.

○ **언어자원 측면**에서는 기존에 구축된 언어자원들의 검증/수정/보완이 필요하며, 국어정보화 기술 수준을 한 단계 도약시키기 위해서 세계적인 수준에 비해 많이 뒤떨어져 있는 **의미자원(의미주석 말뭉치, 어휘의미망)의 구축**이 필요하다. 또한 문어체 뿐만 아니라 구어체, SMS, 인터넷 문서 등 전반적인 **언어생활의 변화를 모니터할 수 있는 다양한 장르**의 언어자원을 꾸준하게 구축할 필요가 있다.

○ **국어자원 활용 측면**에서는 많은 예산을 투자해서 구축한 각종 언어자원들을 모든 국민들이 쉽게 접근하고 활용할 수 있도록 하는 환경이 제공

되어야 한다. 이를 위해 차세대 웹(semantic web) 기반의 **통합 국어자원 관리 및 지원 시스템(한국어 IT HUB)**을 구축하고, 언어자원과 이들의 활용을 위한 각종 언어처리지원 프로그램을 제공할 필요가 있다.

○ 21세기 지식정보화 사회에서는 전문용어가 **지식의 표준화를 위한 국가적 산업인프라**로 인식되고 있다. 이를 위해 정확한 개념에 바탕을 둔 전문용어의 사용으로 국가 지식관리를 **효율화·표준화**하고, 국민 간의 정보 전달을 원활히 하기 위한 **생활 전문용어의 정비 및 대 국민 서비스**가 필요하다.

○ 국어정보화 사업으로 도출된 여러 결과물을 다양한 방안으로 실용화하고 기업에서 이를 산업화할 수 있도록 국어정보화 사업 개발 단계에서부터 **기업이 참여할 수 있는 방안**을 모색할 필요가 있다.

4. 연구내용 및 범위

○ 21세기 세종계획 수립 후의 국어 정보화 환경 분석

- 국내외 정보 환경 변화 분석
- 국내외 언어 처리 기술 및 연구 동향 분석
- 국내 국어 정보화 현황 조사 및 한계점 분석

○ 21세기 세종계획 결과물에 대한 분석

- 21세기 세종계획을 비롯한 기존의 국어정보화 사업을 통해 구축된 자원의 응용적, 활용적 관점에서의 분석
- 산업화를 위해 추가적으로 요구되는 국어정보화 자원의 양과 종류, 형태 분석

○ 국어정보화 2단계에서 중점 추진할 사업으로 다음 세 분야를 선정하였음.

(1) 지식사회를 선도하는 한국어 분야 : 국어자원의 활용

- 한국어 정보처리 인프라 구축을 위한 실용 언어 자원 구축
- 언어 산업과 연계한 언어 인프라 구축
- 국어 자원의 IT 활용을 위한 공유 체계 구축

(2) 미래를 준비하는 한국어 분야 : 의미분석 자원 구축

- 다양한 의미분석 자원 말뭉치 구축
- 호환성을 갖춘 한국어 어휘의미망 구축
- 의미분석 자원 간의 연계/연동 시스템 구축

(3) 세계와 소통하는 한국어 분야 : 전문용어 및 언어자원 표준화 분야

- 전문지식의 보편화와 자생적 융합을 위한 전문용어 표준화 체계 구축
- 국어 언어자원 표준 구축 및 생활 전문용어 구축
- 언어처리 기술 평가를 위한 표준화된 평가 세트 구축

5. 연구결과의 활용

- 본 기획 연구 결과로 도출된 중점 사업은 2010년부터 2014년까지 5년간 단계별로 추진
- 2단계 사업 결과물의 대국민 서비스
- 2단계 사업 결과물의 산업계 활용

차 례

<표 차례>

<그림 차례>

I. 서론

1. 사업 목적

21세기 세종계획은 "우리말과 우리글을 바탕으로 하는 정보사회 건설"을 위해 (1) 세계수준의 국어 기초 언어 자료베이스 구축을 통한 우리말 정보화, (2) 표준화된 전자사전 구축을 통한 우리말 체계화, (3) 한민족 언어 정보화를 통한 우리말 세계화를 사업 목적으로 하여 1998년부터 2007년까지 10년간 수행되었다.

그러나 그동안 정보통신 기술의 급속한 발전과 진화로 지식 사회로 진입이 가속화됨에 따라, 21세기 세종계획을 기획할 당시에 예측하지 못한 **국내외적인 여러 가지 변화를 반영하고, 10년간 진행되어온 21세기 세종계획의 결과물을 적극적으로 활용하여 산업화할 수 있는 방안**이 모색되어야 한다.

특히, 최근에는 개방과 공유, 집단 지성을 모토로 하는 WEB 2.0 방식으로 컴퓨팅의 환경이 바뀌고 있어, 이러한 환경변화에 따라 향후 5년 후의 산업계 및 학계에서 요구되는 국어정보화를 예측하는 사업을 기획할 필요가 있다.

본 사업은 이러한 국내외적인 정보화/지식 사회에 부응하는 국어정보화 2단계 사업을 계획하는 것을 목적으로, 구체적으로 지난 10년간 수행된 21세기 세종계획을 비롯해 기존의 국어정보화 사업을 통해 구축된 자원의 응용적·활용적 관점에서의 분석하고, 그동안 변화된 국어 정보 환경을 면밀히 조사하여, 이를 바탕으로 2010년부터 2014년까지 5년간 국가사업으로 추진할 국어정보화의 중장기 사업 계획을 수립하는 것을 목적으로 한다.

2. 사업의 배경 및 필요성

2.1. 사업의 배경

○ **21세기 세종계획**은 "우리말과 우리글을 바탕으로 하는 정보사회 건설"을 위해 세계수준의 국어 기초 언어 자료베이스 구축을 통한 우리말 정보화, 표준화된 전자사전 구축을 통한 우리말 체계화, 한민족 언어 정보화를 통한 우리말 세계화를 사업 목적으로 하여 1998년부터 2007년까지 총 15,066백만 원의 예산이 투자되었다.

○ 이러한 **21세기 세종계획** 결과물로는 크게 현대국어 기초 말뭉치 구축, 국어 특수 자료 구축, 전자 사전, 전문 용어, 한민족 언어 정보화 DB 구축 등으로 **세계적인 수준의 언어자원이 구축**되었으며, 문자 코드 표준화 연구, 국어 정보처리 표준화 등의 **표준화 작업**이 이루어졌다. 또한, 한민족 언어 정보화를 위한 다양한 검색 프로그램 및 말뭉치 및 전자사전을 구축·활용하기 위한 **다양한 소프트웨어 프로그램**(용례추출기, 형태소분석기, 글잡이, 한마루 등)이 개발되어, 학계 및 정보 산업계 등에 제공함으로써 인력양성 및 국어정보화 차원을 한 단계 끌어올렸다.

○ 그러나, 최근 정보통신 기술의 급속한 발전과 환경 변화로 인해, 국어정보화에 대한 요구사항도 다양해지고 있어 이를 반영할 수 있는 2단계 국어정보화 사업이 필요한 시점이다.

○ 지난 10년간 국어정보화 관련 국내외적인 환경변화는 다음 <표 1>과 같이 요약할 수 있다.

<표 1> 국어정보화 국내외 환경변화

	21세기 세종계획 기획 (1997년)	국어정보화 2단계 계획 (2008년)
언어자원/말뭉치	• KAIST/SERI : 7,600만 • 고려대 : 5,000만 • 연세대 : 4,500만 • 울산대 : 300만 • 국립국어원 : 6,100만(특수) • 단국대 : 옛문헌(15-19세기)	• 현대국어기초말뭉치 • 특수자료(구어,병렬,북한,역사) • 한민족언어정보화 • 문자코드표준화 • 글꼴개발
전자사전	• human readable 형태	• 세종전자사전(NLP용)
언어처리	• 형태소분석(80% 정확률) • 구문분석(초기) • 기계번역시스템(영한/한영/일한: 시제품) • 정보검색(연구,야후) • 의미분석(없음)	• 형태소분석(99% 정확률) • 구문분석(80% 정확률) • 기계번역시스템(영한/한영/일한/한일/중한/한중,특허분야,기술문서분야) • 정보검색(연구,야후) • 의미분석(85% 정확률)
전문용어	• 미비	• 200만(15개 분야)
컴퓨팅 환경	• 초기인터넷/keyword 검색	• WEB2.0/의미기반 검색
표준화	• 미비	• ISO 표준활동 참여

2.2. 사업의 필요성

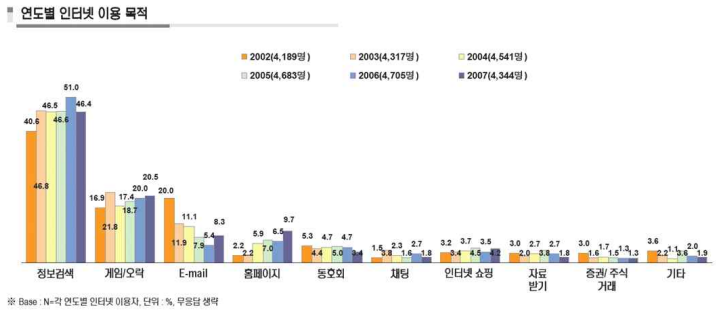
○ 21세기 세종계획이 기획될 당시와 종료된 후 후속적인 국어정보화에 대한 다양한 요구사항은 <표 2>로 요약될 수 있다.

<표 2> 국어정보화에 대한 요구사항

	21세기 세종계획(1997년)	국어정보화 2단계 계획(2008년)
기술환경 및 사회적 요구사항	• 학술적 활용을 초석 단계 • 점진적 활용 및 발전 단계 (철자교정, 기계번역 등 한정된 활용 영역)	• 응용영역 확대를 위한 활용단계 • 급속한 변화와 확산에 대비한 단계 • 사용자 중심의 응용 및 활용으로 서비스 예측이 힘듦 • 인터넷 사용자 증가 및 구어체 중심의 의사교환과 text mining에 대한 요구 증대
언어자원 구축수준 요구사항	• 활용할 언어자원 부족 • 문어체 중심 • 표준 태그 set 부재	• 서로의 협력으로 구체화 단계 (체계화) • 구어체 영역으로 확대 • 개념망으로 발전
IT 분야 활용기술 요구사항	• 형태소 분석 중심 • 단순 검색	• 의미 분석 단계로 발전 • 시맨틱 웹 • 정보분석의 요구 증대 (information intelligence)
표준화수준 요구사항	• 국내용	• 국제표준과 일치
전문가 개발환경 요구사항	• off-line 으로 체계적이지 못했음	• 협업단계로 발전

2.2.1. 기술 환경 및 사회적 요구사항

○ 인터넷 사용자들의 인터넷 이용 목적은 <그림 1>과 같이 2002년도부터 정보검색의 요구가 1위를 차지하고 있음.



<그림 1> 연도별 인터넷 이용 목적
(출처 : KOBACO, "2007 소비자 행태 조사")

○ 국내 포털 사업자들은 서비스의 경쟁 우위 확보를 위해 검색 분야에 대한 R&D 투자를 늘리고, 새로운 검색 서비스를 발굴하고자 하는 노력을 경주하고 있음.

○ 모든 포털들이 언어 자원의 구축과 운영에 자체적으로 많은 예산을 투입하고 있는 상황에서, 중복 투자를 방지하고 검색 서비스의 확장을 위해서는 **신조어, 전문용어 등의 언어 자원을 지속적으로 확충**해야 함. 또한 검색 서비스의 품질 개선을 위해 공신력 있는 **평가 세트가 없어 자체적으로 테스트 환경을 구축**하여 품질 개선 활동을 하고 있음.

○ 한편으로 최근 구글이 한국어를 포함한 다국어 번역 서비스를 개시함에 따라 정보검색뿐만 아니라 기계번역 분야까지 포털 사업자의 경쟁이 확대될 전망임.

○ 해외 포털들이 한국 진입이 가속화됨에 따라 이에 대한 가장 큰 장벽이 되고 있는 **언어처리기술에 대해 경쟁 우위를 점할 필요성**이 증가하고 있음.

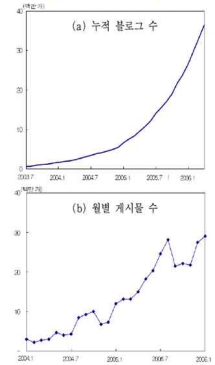
○ 시장 환경의 변화에 따라 포털 사업자들의 국어자원의 활용 필요성이 점차 더 높아지고 있음.

○ 블로그(blog)를 중심으로 한 UCC의 활성화에 따라 비정형 문서에 대한 분석 기술을 중심으로 한 언어처리에 대한 중요성이 증가하고 있음.

○ 아래 <그림 2>에서 보여 주는 것과 같이 국내 블로그 및 포스트는 2007년에 비해 2008년 9월 현재 2배 이상 성장한 것으로 파악되며 이에 따라 급증하고 있는 블로그 포스트를 기반으로 한 검색 기술, 분류 기술,

필터링 기술 등이 요구되며, 이의 기반이 되는 **언어자원의 구축과 응용 기술 개발**이 요구되고 있음.

세계 블로그 증가 추이



출처: [OECD, "Information Technology Outlook", 2006] ※세계 추이에 국내 데이터는 제외되어 있음

국내 15개 주요 사이트의 블로그 및 포스트 수(2007년 7월)

사이트	블로그 수		포스트 수		평균 포스트 수
naver.com	2,925,174	57.6%	57,787,502	59.4%	19.8
daum.net	669,412	13.2%	18,509,316	19.0%	27.7
nate.com	402,750	7.9%	2,092,885	2.2%	5.2
yahoo.com	345,220	6.8%	5,371,066	5.5%	15.6
paran.com	336,973	6.6%	2,797,578	2.9%	8.3
empas.com	170,893	3.4%	4,391,599	4.5%	25.7
egloos.com	81,746	1.6%	2,588,077	2.7%	31.7
tattertools	49,441	1.0%	1,742,428	1.8%	35.2
joins.com	30,512	0.6%	866,963	0.9%	28.4
chosun.com	43,875	0.9%	442,504	0.5%	10.1
news.go.kr	3,614	0.1%	262,977	0.3%	72.8
dreamwiz.com	11,964	0.2%	240,275	0.2%	20.1
ohmynews.com	2,976	0.1%	58,350	0.1%	19.6
jinbo.net	1,814	0.0%	25,843	0.0%	14.2
mediamob.co.kr	1,023	0.0%	37,530	0.0%	36.7

출처: Yahoo! Korea 검색 프로파일

<그림 2> 국내 블로그 및 포스트 증가 추이

2.2.2. 언어자원 구축 수준 요구사항

○ 말뭉치는 언어 정보화의 가장 기초적인 자료이므로, 적절한 양의 정확한 말뭉치를 구축하여야 한다. 정확률이 전제되지 않으면 말뭉치의 활용을 기대할 수 없다. 21세기 세종계획에서 방대한 양의 말뭉치를 구축했음에도 불구하고 그 활용도가 그렇게 높지 않은 것은, 말뭉치의 **응용 분야의 요구에 따른 다양한 형태**를 구축하지 못하여 응용분야의 개발자마다 이를 재가공해야 하는 문제가 있음.

○ 원시말뭉치는 **언어생활의 변화를 모니터**링 할 수 있도록 꾸준하게 구축되어야 한다. 특히, 인터넷과 휴대폰이 생활화되면서, 우리 국민이 많이 활용하고 기존 언어와 다른 양상을 보이는 SMS, 인터넷 문서 등의 **다양한 장르**에 대한 원시 말뭉치의 구축도 필요함.

○ 형태소 분석 말뭉치는 정확률을 99% 이상이 되도록 개선해야 한다. 최근 한국어 형태소 분석기 및 자동 태거의 통합 정확도가 97% 이상인 시스템들이 개발되어 있기 때문에 이러한 시스템을 활용하면 수작업을 최

소화하면서 대규모의 형태소 분석 말뭉치를 구축할 수 있음.

○ 구문 분석 말뭉치의 정확률 역시 일정 수준 이상 되도록 개선해야 하며, 다양한 분석 요구에 맞도록 수정되어야 한다. 그동안 한국어 구문 분석에 대한 연구가 활성화되지 못한 가장 중요한 이유는 **다양한 분석 형태에 적절한 구문분석 말뭉치, 하위범주화 정보, 시소러스, 어휘의미망** 등의 언어자원이 제공되지 못했기 때문이다. 21세기 세종계획에서 구축한 약 82만 어절에 대한 구문 분석 말뭉치는 양적으로는 적지 않으나, 다양한 분석 요구에 맞도록 수정되어야 함.

○ 형태의미 분석 말뭉치는 의미부류에 대한 체계적인 검토를 바탕으로, **전자사전의 의미부류와 연계**되도록 재정리되어야 한다. 특히 **전자사전과 표준국어대사전이 연계**되어야 다양한 언어처리 시스템에 활용될 수 있음.

○ **어휘의미망**은 21세기 세종계획에 포함되어 있지 않았지만, 한국어 언어처리 응용 분야에서 가장 필요로 하는 언어자원 중 하나이므로, 모든 연구자들이 접근하여 활용할 수 있는 신뢰성 있는 어휘의미망을 구축하여야 함.

2.2.3. IT 분야 활용기술 요구사항

○ 언어처리 기술이란 일상생활에서 쓰는 자연스러운 언어를 사용하여 정보기기를 제어하거나 정보서비스를 받을 수 있도록, 말과 글로 표현되는 언어를 단순 저장, 변환하는 수준을 넘어 그 안에 포함된 정보를 추출, 가공, 활용하는 기술로서 인간처럼 말하고, 듣고, 보고, 이해하게 하는 기술 등을 포함함.

○ 언어처리 기반기술은 입력된 구어체 및 문어체 텍스트의 내용에서 문장에 나타나는 단어의 품사를 분석하는 형태소 분석 기술, 문장의 구조를 분석하는 구문 분석 기술, 문장의 의미를 파악하는 **의미 분석 기술**, 대화의 의미를 파악하는 **담화 분석 기술**이 포함됨.

○ 검색 기술(텍스트 마이닝)은 지식관리 및 지식검색 등 타 소프트웨어 분야의 핵심 및 부가기능을 담당하여 타 분야에 대한 파급효과는 매우 큼.

○ 특히 차세대 웹(semantic web)은 인터넷의 발달과 함께 수요가 증가하

고 있으며 사용자의 편의성에 초점을 맞춘 자연어 질의와 **의미기반의 정보검색기술**이 각광받고 있으며, 사용자의 질문에 답변을 제공할 수 있는 **포털 지식제공 서비스**가 대두되고 있음.

○ 의미기반 정보검색을 위해서는 엔티티 추출, 의미 중의성 해소, 개념 추출 등의 요소 기술이 개발되어야 하며, 이를 위해서는 의미분석 자원**(의미주석 말뭉치, 어휘의미망)**이 필요함.

2.2.4. 표준화 수준 요구사항

○ 지금까지는 언어 자원이나 언어 정보 처리 기반 기술이 개발자 자신만이 쓸 수 있는 모습이어서 연구자들 간에 호환이나 공유에 의한 재사용이 대부분 불가능하였고, 이러한 상황은 쓸데없는 중복 투자를 야기하며 언어정보 처리 분야의 발전에 한계를 맞게 하였음.

○ 이러한 문제점들을 해결하기 위한 가장 근본적인 해결책으로서 90년대 초부터 연구자들 간에 언어 자원의 표준화에 대한 필요성이 대두되었음.

○ 언어처리 응용 기반 기술과 관련해 다양한 평가가 이루어지고 있음: 형태소 분석 API 기술(MorphOympic), 구문 분석 API 기술(유럽의 프랑스, 이태리, 독일, 스위스, 영국 등이 국제 공동 연구로 수행한 다국어 언어처리 평가 세트 마련을 위한 프로젝트인 TSNLP가 수행됨), 의미 분석 API 기술 평가 대회(SENSEVAL). 따라서, 산업체 및 학계에서 개발한 언어처리 기술을 평가하기 위한 표준화된 다양한 **평가 세트가** 구축되어야 함.

○ 자동 번역 및 정보검색/텍스트마이닝 엔진 평가는 관련 제품의 성능 향상을 통한 산업 활성화에 큰 기여를 할 수 있으므로, **언어처리 기반 기술 평가를 위한 평가 세트를 표준화함으로써 객관적인 평가체계**를 마련할 필요가 있음.

○ 한국어에 고유한 영역의 기술에 대하여 **국내 표준화**를 조기 추진하여, 외국어를 대상으로 개발된 기술이나 방식을 기계적으로 적용함으로써 야기되는 문제를 사전에 방지함.

○ 현재 전문용어 표현 양식에 국한하여 참여하고 있는 국외 표준화 활동을 확대하여 언어처리응용 기반기술인 언어분석 API 표준화와 언어처리

응용엔진의 평가 체계 구축의 표준화에도 국내 표준화 포럼을 통해 적극 참여함.

○ 인터넷 정보가 급속도로 증가하면서 사용자들의 정보 검색에 대한 성능, 그리고 이의 개선에 대한 요구가 증가함에도 불구하고 이를 실질적으로 평가하기 위한 표준화된 기준과 기반이 마련되지 않고 있으며, 이에 대한 연구가 시급한 실정임.

○ 전문용어는 인터넷 검색의 색인 혹은 키워드로 쓰인다. **전문용어의 표준화**가 되지 않으면 한 개의 개념에 대하여 여러 형태의 단어를 쓰게 되어, 검색이 충분히 이루어지지 않으며, 관련된 지식과 정보가 통합되지 않음.

○ 21세기의 IT 기반 지식정보화 사회에서는 전문용어가 **지식의 표준화를 위한 국가적 산업인프라**로 인식되고 있음.

○ 지식정보 유통의 다원화로 인해 외국어 특히 영어 용어에 대한 수용 방식에 있어서 고유어, 한자어, 원어 차용 등 용어사용의 개인화로 인해 다양한 용어 사용 방식이 난립하고 있음..

○ 정확한 개념에 바탕을 둔 전문용어의 사용으로 국가 지식관리의 효율화와 국민 간의 정보 전달을 원활히 하기 위해 필요함.

○ 특히 인터넷을 통해 쏟아지는 각종 전문분야 정보의 대중화로 인한 생활 전문용어 정비 및 대 국민 정보서비스의 필요성이 증대되고 있음.

○ 전문용어의 정비 및 표준화를 통하여 정보의 효율적 유통 체계가 구축되어야, 정보산업 응용기술 개발에 기여할 수 있으며 국가 산업의 인프라로 자리매김할 수 있음.

○ 정보사회로의 조기 진입을 위한 대단위 정보처리 기술은 대용량 한국어 정보처리 기술 및 각 전문 분야마다 독특성을 반영하여야 함.

○ 전문적 국어정보처리에 대한 품질 관리는 전문분야별 전문용어의 확립으로 향후 안정된 한국어 정보처리 기술의 확보와 산업화에 효과가 있음.

2.2.5. 전문가 개발환경의 요구사항

○ 최근 인터넷에서는 개방과 공유, 집단 지성을 모토로 하는 web 2.0 방식에 의해 인간의 지식을 구축하거나, 새로운 정보에 대해서 사용자들의 평가를

달아주는 노력이 활발히 이루어지고 있음.

○ 가장 대표적인 사례는 백과사전을 구성하기 위해 집단지성을 활용하는 위키피디아(Wikipedia)와 오프소스 개발 프로젝트를 통해 오픈소스를 보급하고 있는 SourceForge로 개발기간 단축 및 콘텐츠의 양질화, 콘텐츠 이용의 확대를 위해 집단지성을 활용하는 추세임.

○ 따라서, 언어자원의 효과적인 구축/활용과 언어처리 기반 기술 개발 환경을 **공유, 참여, 협업 방식이 필요**하며 이를 위해 **WEB 2.0 방식을 적용**할 필요가 있음.

2.3. 국어정보화 2단계 중점 추진 사업

이상과 같은 여러 요구사항을 충족시키기 위한 향후 5년간 우선 진행될 2단계 국어정보화 사업의 중점 추진 사업으로 다음의 세 분야를 선정하였다.

(1) 지식사회를 선도하는 한국어 분야 : 국어자원의 활용

○ 한국어 정보처리 인프라 구축을 위한 실용 언어 자원 구축
○ 언어 산업과 연계한 언어 인프라 구축
○ 국어 자원의 IT 활용을 위한 공유 체계 구축

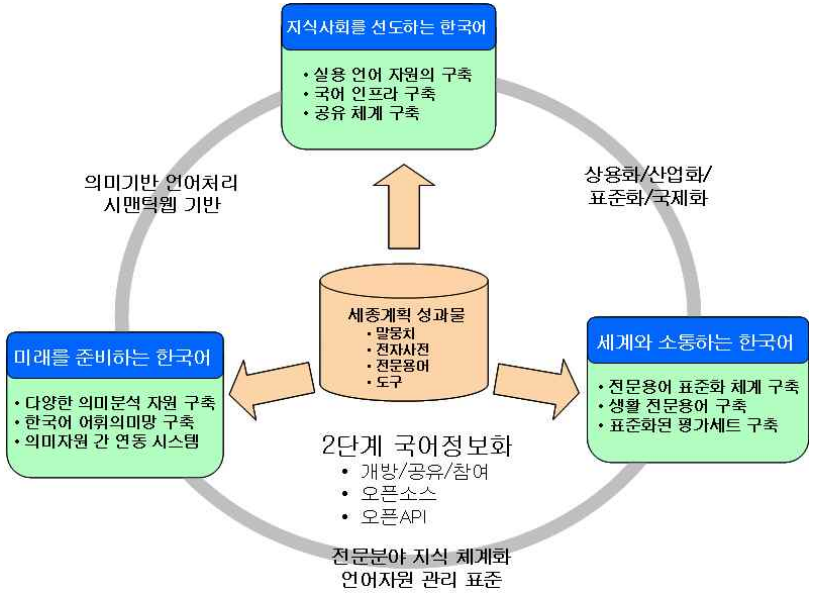
(2) 미래를 준비하는 한국어 분야 : 의미분석 자원 구축

○ 다양한 의미분석 자원 말뭉치 구축
○ 호환성을 갖춘 한국어 어휘의미망 구축
○ 의미분석 자원 간의 연계/연동 시스템 구축

(3) 세계와 소통하는 한국어 분야 : 전문용어 및 언어자원 표준화 분야

○ 전문지식의 보편화와 자생적 융합을 위한 전문용어 표준화 체계 구축
○ 국어 언어자원관리 표준 구축 및 생활 전문용어 구축
○ 언어처리 기술 평가를 위한 표준화된 평가 세트 구축

위 세 분야의 사업 간에는 연계성과 국어정보화 2단계 사업의 비전은 다음 <그림 3>과 같다.



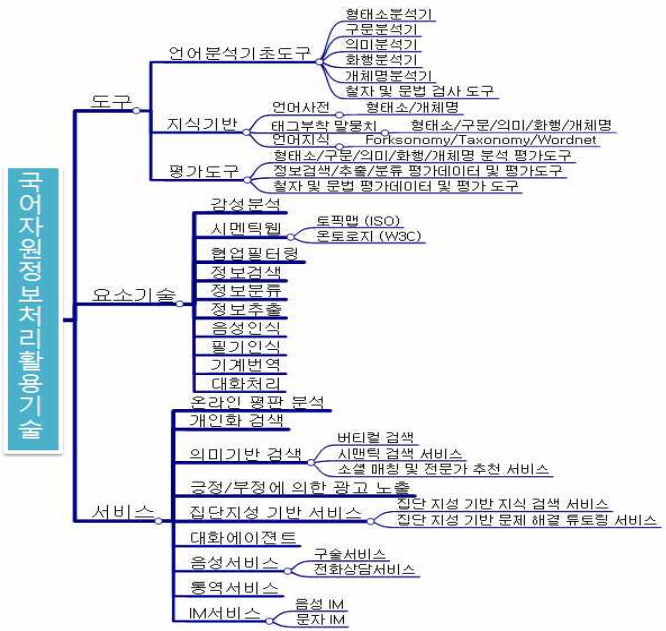
<그림 3> 국어정보화 2단계 사업의 비전

II. 추진 사업 분야별 국내외 국어정보화 현황

1. 언어자원 활용 분야

1.1. 언어자원 활용 현황

○ <그림 4>는 언어 자원을 활용하는 현재의 기술을 도구, 요소기술, 서비스로 나누어 도식화한 것임. 서비스는 요소기술을 기반으로 하며 요소기술은 도구를 사용해서 이루어짐.



<그림 4> 언어자원을 IT에서 활용하는 기술

○ 언어 자원을 IT에서 활용하는 방법에는 정보검색, 분류, 추출 등의 정보처리, 음성이나 문자 인식 등의 언어인식, 기계번역, 대화처리 등 다양한 분야가 있으나 현재 주류를 이루는 분야는 인터넷에서 포털로 대변되는 정보검색 분야임.

○ <표 3>은 코리안 클릭의 2008년 8월 페이지뷰 순위로 인터넷을 이용하는 인구의 대부분은 포털이나 쇼핑몰, 블로그 등의 정보처리 분야 서비스를 이용하고 있는 것을 볼 수 있기 때문에 IT 분야에서 언어자원의 활용성을 높이려면 이와 관련된 기술 개발이나 연구가 많이 있어야 함.

<표 3> 2008년 8월 국내 도메인별 페이지뷰 순위

순위	도메인	순방문자 (*1000)	도달률 (%)	순위	도메인	순방문자 (*1000)	도달률 (%)
1	www.naver.com	31,152	96.59	▲6	www.auction.co.kr	18,410	57.08
2	www.daum.net	29,694	92.07	▲7	www.yahoo.co.kr	18,010	55.84
3	www.cyworld.com	24,310	75.38	▼8	www.gmarket.co.kr	17,586	54.53
4	www.nate.com	22,631	70.17	▲9	www.tistory.com	17,537	54.38
5	www.empas.com	19,485	60.42	▼10	www.paran.com	15,780	48.93

○ 한국어가 지식사회를 선도해 나아가려면 한국어로 된 정보들을 정확하게 수집하고, 그 언어현상을 쉽게 확인할 수 있으며, 이의 적절성을 평가하고, 각 컴포넌트가 바른 한국어의 언어현상을 측정할 수 있는 기반 환경이 마련되어야 함.

1.1.1. 언어자원 구축 현황(21세기 세종계획 결과물)

○ 21세기 세종계획 결과물은 크게 국어 기초 자료 구축, 국어 특수 자료 구축, 전자 사전, 전문 용어, 한민족 언어 정보화 DB 구축, 문자 코드 표준화 연구, 국어 정보처리 표준화, 핵심 소프트웨어 개발 등으로 구분되어 있음.

○ 다음의 표들은 21세기 세종계획에서 구축된 언어자원을 나타내고 있으며, <표 4>는 국어기초자료 구축현황, <표 5>는 국어특수자료 구축 현황, 그리고 <표 6>은 전자사전 구축 현황, <표 7>는 전문용어 구축 현황, <표 8>는 한민족 언어정보화 DB 구축현황임.

<표 4> 21세기 세종계획 국어기초자료 구축 현황

구 분		구축 어절수
원시 말뭉치	문어	60,558,573
	구어	3,340,839
분석 말뭉치	형태 분석	15,226,186
	형태의미 분석	12,642,725
	구문 분석	826,127

<표 5> 21세기 세종계획 국어특수자료 구축 현황

구 분		구축 어절수
구비 문학	원시	2,363,967
구어 전사	원시	3,671,322
	형태 분석	1,008,681
한영 병렬	원시	4,753,522
	형태 분석	1,009,715
한일 병렬	원시	1,101,878
	형태 분석	299,615
북한 해외	원시	9,505,616
	형태 분석	1,622,337
역사	원시	5,650,834
	형태 분석	883,120
한·중·불·러	원시	150,853
전문용어	원시	1,000,067

<표 6> 21세기 세종계획 전자사전 구축 현황

구분	기본사전 어휘수	상세사전 어휘수
체언	121,500	50,200
용언	52,200	27,150
고유명사	108,600	22,000
조사어미	2,650	2,650
부사	42,224	4,500
연어	15,000	9,000
관용표현	5,000	5,000
특수어	60,000	30,000
복합명사구	14,200	5,600
합계	433,774	169,900

<표 7> 전문용어 구축 현황

구분	말뭉치 (어절수)	한영(일)대응DB (어휘수)
경제학	1,000,000	10,000
물리학	500,000	15,000
화학	1,000,000	15,000
생물학	1,000,000	15,000
의학	1,000,000	15,000
수학	1,000,000	10,000
전산학	1,000,000	30,000
전자전기공학		15,000
기계공학		15,000
산업공학		5,000
화학공학		15,000
재료공학		5,000
환경공학		5,000
건축공학		5,000
토목공학		5,000
합계	6,500,000	180,000

<표 8> 21세기 세종계획 한민족 언어정보화 DB 구축 현황

구분	어휘 수
한글 맞춤법	11,600
남북한언어비교사전	10,000
표준어검색	4,504
외래어	36,437
로마자	11,296
방언	40,151
남북한 이질화된 언어	3,076
어휘 역사	5,129
한국 전통문화 어휘	300

1.1.2. 언어자원 활용 현황

○ <표 4>에서 알 수 있듯이, 원시말뭉치는 문어와 구어를 합쳐 약 6천 3백만 어절, 형태소 분석 말뭉치는 1천 5백만 어절, 형태의미 분석 말뭉치는 1천 2백만 어절, 구문 분석 말뭉치는 82만 어절이 구축되었음. 이러한 말뭉치는 양적으로는 선진국에 비해서도 뒤쳐지지 않을 정도이나 이러한 말뭉치가 지금보다 더 자연언어처리 연구자나 시스템 개발자들에게 활용될 여건을 조성할 필요가 있음.

○ 사업 시작 당시와 달라진 저작권 환경으로 원시말뭉치의 저작권 해결이 아직 진행 중에 있으며, 또한 형태소 분석 말뭉치와 구문 분석 말뭉치의 완성도를 높여 다양한 응용 분야의 구체적인 요구사항을 충족시킬 수 있게 보완하는 것이 바람직함.

○ 형태의미 분석 말뭉치와 세종전자사전의 보다 적극적인 활용을 위해서는 의미 부류에 대한 체계적인 검토를 통해, 두 자료 간의 연계성을 높이는 것이 의미적 중의성 해결 분야의 활용성을 높이는 방안임.

1.2. 포털 및 실생활 응용

○ 국내 포털 사업의 발전 과정은 검색서비스의 발전과 밀접한 관련을 가짐.

○ 1990년대에는 검색엔진이라는 이름으로 야후, 심마니, 라이코스 등이 서비스를 제공했으며, 이 당시의 주요 서비스는 "디렉터리 검색" 및 "웹페이지 검색"이었음.

○ 메일과 카페를 핵심서비스로 한 다음(daum)의 약진이 있었으나 "지식검색"을 핵심서비스로 한 네이버에 선두 자리를 내어주고 오늘에 이르고 있음.

<그림 5> 국내 포털 사업의 발전 과정

○ 최근 국내 모든 포털들은 검색을 다시 경쟁 포인트로서 내세우는 추세임.

○ 네이버의 경우,

- 한게임의 일본 진출 실패를 만회하기 위해 2010년 초에 일본에서 검색서비스를 론칭할 예정임.
- 이와 병행하여 한국어 검색 서비스로서는 분야별 전문 영역을 확대하고 있음. 아울러 이미 인물, 영화, 게임, 자동차 등의 분야별 전문 검색을 추진하고 있음.

○ 다음(daum)의 경우,

- 동영상 UCC 이후 카페 검색을 강화하고 있으며, 추가로 10여 개 섹션의 분야별 전문 검색을 출시하고자 준비 중임.
- 티스토리를 M&A함으로써 다양한 전문 블로거들의 양질의 콘텐츠를 확보하여 이를 바탕으로 커뮤니티 뿐만 아니라 검색 서비스에도 활용할 방안을 모색 중임.

○ SK커뮤니케이션즈 역시 검색 기술의 강화를 위해 엠파스를 M&A한 바 있으며, 올해 들어

- 미니홈피를 중심으로 한 커뮤니티 포털인 싸이월드를 종합포털 형태로 개편하였으며, 뉴스 및 통합검색을 강화하고 있음.
- 특히 미니홈피 콘텐츠를 검색하고 인명 검색 기능을 제공함으로써 차별화를 꾀하고 있음.

○ 야후의 경우, 야후거기의 인지도를 바탕으로 위성지도 등을 강화하며 지역검색의 특화를 노리고 있음.

○ 파란의 경우,

- 탄탄한 지역정보/지도 콘텐츠를 바탕으로 지역검색을 강화하고 있음.
- 연초부터 주제 집중 검색이라는 분야별 전문 검색을 시도하고 있어 게임, 취업, 재테크, 맛집 등 다양한 주제별로 아웃링크 방식으로 깊이 있는 검색 서비스를 제공하고 있음.

○ 종합 포털들은 검색 서비스의 경쟁에서 이기지 못하면 사업의 유지가 어렵다는 판단 하에, 경쟁력 있는 검색 콘텐츠를 확보하고, 차별화된 검색 기능을 제공하고자 노력하고 있음.

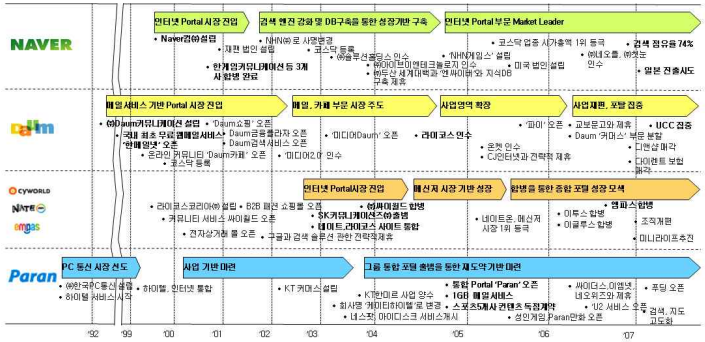
○ 2005년 이후 UCC의 발전, 블로고스피어의 확대 등이 인터넷 시장의 큰 흐름에 따라 포털에서도 동영상 UCC 및 블로그 서비스를 강화하고 있는 추세임.

○ UCC 및 블로그 서비스의 접근성을 높이고 콘텐츠를 원활히 소비하게 하기 위해서는 언어처리 기술들이 요구되고 있음.

○ 한국의 블로그는 타인의 글이나 뉴스 콘텐츠를 스크랩하는 경우가 많은데, 검색 시 이러한 포스트들이 상당 부분 중복되어 제시됨. 주로 펌질에 의해 전파되는 동영상 UCC도 비슷한 양상을 보이고 있으며, 포털 사업자들은 이러한 UCC들의 중복을 제거하고 필터링하는 기술의 개발

을 추진하고 있음.

○ 저작권의 강화, 청소년보호 강화, 인터넷 문화의 정화 등의 사회/제도적 이슈와 맞물려 UCC에 대한 모니터링 노력도 점차 증가하고 있음.



<그림 6> 국내 포털의 주요 서비스 론칭 과정

2. 의미분석 자원 구축 분야

2.1. 의미관련 국내외 연구개발 현황

2.1.1. 국외 연구개발 현황

○ 자연언어처리기술은 점차 형태, 통사기반에서 의미기반으로 변화하고 있음. 이는 좀 더 지능적인 언어능력을 갖춘 시스템에 대한 사용자들의 수요와 관련이 있음.

○ 예를 들어, 인터넷 검색 분야에서는 종래의 단순 웹 기반 질의어 탐색에서 문장의 의미와 사용자의 의도를 파악해야 하는 온톨로지 기반 시맨틱웹 검색으로 변화하고 있음.

○ 더 나아가 문서의 논조(sentiment)를 파악하여 특정 테마에 대한 사용자의 반응을 자동으로 분석하는 시스템까지 개발되고 있음. (IBM 연구보고서 참조: http://www.trl.ibm.com/projects/textmining/takmi/sentiment_analysis_e.htm)

○ 기계번역 및 자동통역 분야에서는 기존의 문장단위의 한계를 뛰어넘는 텍스트 번역 및 구어체 문장번역을 위한 연구가 진행되고 있음. 문장 단위를 뛰어넘는 텍스트의 기계번역을 위해서는 문장의 통사분석은 물론 의미분석이 이루어져야 함.

○ 최근 IBM 및 일본 문부성 등에 의해 향후 가장 유망한 IT분야의 하나로 손꼽힌 자동통역분야는 기존의 기계번역 시스템과는 달리 구어체 문장을 처리해야 함.

○ 이에 미국 국방부 연구소인 DARPA에서는 군사분야 자동통역시스템의 개발을 착수함. 유럽연합에서도 매트릭스라고 하는 유럽어 간 자동통역 시스템 개발을 착수함. 일본에서는 국책연구소인 NICT와 ATR 등을 중심으로 일-영, 일-중 자동통역시스템 개발이 진행되고 있음.

○ 구어체 문장번역은 문장요소의 생략, 대용어, 은유, 대화참여자의 의도 파악 등 여러 가지 의미적인 요소의 처리 없이는 불가능함.

○ 이와 같은 의미관련 정보의 구축을 위해서는 의미정보가 풍부하게 부착된 말뭉치가 필수적임.

2.1.2. 국내 연구개발 현황

○ 21세기 세종계획 등을 비롯한 연구프로젝트에서 구축된 언어자원은 주로 형태, 통사정보와 관련된 언어자원임(사전, 말뭉치 등).

○ 21세기 세종계획의 결과물은 언어처리 기반기술(형태소분석, 구조분석 등) 및 기본적인 언어처리 응용시스템(키워드 기반 정보검색, 문장단위 기계번역 시스템, 맞춤법 검사기 등)의 개발을 위해 반드시 필요하나, 이를 넘어서는 고급 응용시스템의 개발을 위해서는 좀 더 풍부한 의미 정보가 필요함.

○ 고급 응용시스템을 개발하려면 언어처리 기반기술의 획기적인 성능향상이 있어야 함.

○ 언어처리 기반기술의 획기적인 성능향상은 의미분석이라는 벽을 넘지 않고서는 불가능함.

○ 현재 국내 형태소분석 기술수준은 약 99% 이상이라고 보고되고 있지만, 태깅(tagging) 정확률을 따지면 95% 미만일 것으로 예상됨.

○ 이러한 태깅 성능의 한계는 대부분 의미를 고려하지 못하는 현 방법론의 한계와 관련이 있음.

○ 이와 유사하게 문장의 구조분석 성능도 현재는 단문단위 구조분석 정확률이 85~90% 정도의 수준인 것으로 보고됨.

○ 단문단위 구조분석 정확률이 95% 이상 올라가지 않으면 텍스트 기계번역이나 자동통역 시스템의 개발은 거의 불가능함.

○ 21세기 세종계획의 결과물 중 일부 의미와 관련된 자원(의미정보부착 말뭉치)이 존재하나, 앞에서 언급한 연구개발을 위해서는 양과 질이 모두 부족한 실정임.

○ 따라서 한국어 관련 언어처리 기반기술의 획기적인 발전 및 고급응용시스템 개발을 위해서는 다양한 의미정보가 부착된 말뭉치가 필요함

○ 또한 시맨틱웹 환경에 성공적으로 적응하려면 통일된 한국어 어휘망의 필요성이 대두되고 있음

○ 이를 위해 기존에 구축된 대표적인 한국어 관련 어휘망을 통합 구축해야 할 필요가 있음.

2.2. 의미분석 자원 구축 현황

2.2.1. 논항정보부착 말뭉치 구축

2.2.1.1. 국외 연구현황

○ 문장의 논항정보, 또는 의미역 정보는 동사를 중심으로 그와 연관된 요소들의 의미적 특성을 규명해 놓은 정보로서, 행위주(Agent), 대상(Theme) 등으로 분류함.

○ 문장의 통사구조적 특성에 못지않게 문장의 논항정보가 문장의 의미적 특성을 파악하는데 매우 중요하다고 보고 있음.

○ 논항정보의 중요성에 대한 연구는 언어학에서 1970년대 Fillmore의 “The case for case"부터 시작하여 언어현상의 이해에 핵심적인 부문으로 간주되어 오고 있음.

○ 전통적으로 이러한 논항정보는 연구자의 직관에 의존하여 파악하고 정

리하여 왔음.

○ 언어자원의 대규모화와 함께 논항정보를 전산적 방법으로 대규모로 파악, 추출하는 문제도 학자들의 관심의 대상이 되고 있음.

○ 특히 구문분석 말뭉치의 성공적인 구축 이후에 논항구조 정보를 담고 있는 말뭉치에 대한 관심이 증대됨.

○ 논항구조정보를 파악하는 방법으로는 동사별 논항구조를 체계적으로 정리하는 방법, 논항구조정보가 부착된 말뭉치를 구축하는 방법, 그리고 원시말뭉치나 구문분석말뭉치에서 논항정보를 자동추출해내는 방법 등이 있음.

○ 동사별 논항구조를 체계적으로 정리하는 연구는 논항구조정보 중심의 어휘 자원을 대규모로 구축하는 작업을 예로 들 수 있음.

○ 논항정보가 부착된 말뭉치 구축은 기존의 통사/구문 분석 말뭉치에 논항정보를 추가해주는 방식이 있음.

○ 논항정보를 원시말뭉치나 구문분석 말뭉치로부터 자동으로 파악하는 연구도 일부 이루어지고 있음.

○ 논항정보 부착 말뭉치 중 현재까지 구축된 대표적인 국외의 말뭉치는 다음과 같음.

(1) Framenet

- http://framenet.icsi.berkeley.edu/
- Berkeley FrameNet은 프레임 의미론에 기반을 둔 온라인 영어 어휘 자원 구축 작업.
- 각 어휘의 하위 의미별로 통사 의미적 결합 가능성 (항가)를 기록하는 것을 목표로 하고 있음.
- 이 작업의 대표적 결과물인 FrameNet는 현재 10,000개 이상의 어휘 단위를 포함하고 있으며, 그 중 6,100개는 필요한 주석처리가 모두 끝나 있고, 825개 이상의 의미 프레임, 그리고 135,000개 이상의 주석처리 된 예문을 포함함.
- 연구자용은 무료, 상업용은 유료 배포

(2) *Propbank*

- http://verbs.colorado.edu/~mpalmer/projects/ace.html
- 구문분석 말뭉치에 논항정보를 추가한 Propbank가 구축되어 있음.

- 대표적인 예로는 Penn TreeBank II의 WSJ 분석한 것이 있음: total propositions: 112,917; total verbs framed: 3,323; total framesets: 4,659; Verbs with multiple framesets: 726; Average framesets per verb: 1.40
- 유료배포

(3) *Penn Korean Treebank*

- http://repository.upenn.edu/cgi/viewcontent.cgi?article=1026&context=ircs_reports
- Penn Korean Treebank에서 두 개의 말뭉치 (각 54,000, 131,000 어절)를 근간으로 한 한국어 Propbank가 구축되어 있음: total verbs framed: 2,749;
- 유료배포

2.2.1.2. 국내 연구현황

○ 일부 대학 연구실 등에서 개별적으로 구축하였거나 구축하고 있을 가능성이 있으나, 외부로 공개된 것은 없음.

○ 가장 근접한 말뭉치로는 ETRI에서 개발해 유료공개 중인 의존관계구문분석, 그리고 세종계획 일환으로 개발된 구문분석말뭉치 등이 있으나 두 경우 모두 의미역이(행동주, 대상. ...) 아닌 문법관계(주어, 목적어, ...) 표지를 붙인 구문분석 말뭉치임.

○ 한국어 논항구조 말뭉치를 이용한 국내의 연구도 드믄 편임. (Han, et al. 2002)

○ 세종전자사전에는 용언의 논항정보가 매우 상세하게 기술되어 있음, 그러나 그러한 정보를 이용한 연구는 아직 거의 없음.

○ 논항구조 자체에 대한 언어학적 연구는 어느 정도 이루어지고 있음.

2.2.2. 개체·시간·공간(ETS) 정보 부착 말뭉치 구축

2.2.2.1. 국외 연구현황

○ 1990년대 말까지 자연언어 의미처리 대상은 이 분야의 기술적인 한계 때문에 주로 어휘(word)와 문장(sentence) 단위에 국한되었기 때문에, 한 개의 텍스트 전체에 기술된 사건 간의 관계나 분산된 다수 텍스트, 다국어 텍스트, 화자 다수가 참여하는 대화문에 나타난 유관 사건 간의 관계를 전혀 파악할 수 없었음.

○ 개체명(named entity) 인식은 1990년대부터 자연언어처리에서 극복해야 할 문제점으로 드러나고 활발한 연구가 진행됨.

○ 이전에 구축한 다양한 언어자원과 통합된 기술력을 바탕으로 하여 2000년대에 들어서면서, 자연언어처리 영역에서 구문 분석과 어휘의미 분석 기술의 완성도가 심화되고, 동시에 지식처리 영역에서 전문 분야 온톨로지(domain specific ontology)가 광범위하게 구축・사용되고 있음.

○ 자연언어로 기술된 분산 텍스트 환경에서, 유관 사건에 대한 유의미한 정보를 추출하고 이를 바탕으로 한 차원 높은 지식을 추론하고자, '누가(who), 언제(when), 어디서(where), 무엇을(what), 하다(do)'와 같이 사건(event)을 구성하는 개체(entity), 시간(time), 공간(space) 정보를 부착함.

○ 자연언어처리와 지식처리를 위한 기술력과 기반 언어자원이 가장 풍부히 구축된 영어(미국)를 중심으로 1990년대 중반부터 소규모로 시작되었고, 서유럽어로 확산되다가, 9/11 사건 이후 아랍어, 중국어 등을 대상으로 삼음.

○ 2000년 초반부터 각각의 ETS 정보를 주석하기 위한 메타언어의 국제표준(International Standard)이 DAML, MITRE, ISO 등에서 활발히 제안되고 있으며, 표준을 통합하는 단계임.

○ 현재는 ETS 정보 주석에 대한 연구를 각각 진행하고 있으나, 향후 통합하여 서비스하기 위한 기반을 마련한 상태이고, LDC(Linguistic Data Consortium)를 통해 주석말뭉치를 배포함.

○ 아직 연구 초기 단계이므로 공간 정보 말뭉치는 아직 구축되지 않았으며, 개체와 시간 정보 주석 말뭉치는 소규모로 구축되고 있음. 평가용 말뭉치를 동시에 제공하여, 말뭉치 구축 초기부터 상이한 기관에서 구축한 말뭉치의 품질을 평가할 수 있음.

(1) *개체 정보 주석 표준안 및 말뭉치*

- ACE(Automatic Content Extraction)
  - http://projects.ldc.upenn.edu/ace/

  - 사건(event)을 구성하는 개체(entity), 관계(relation), 링크(link) 정보 기술
  - 영어, 중국어, 아랍어를 대상으로 guideline 및 주석 말뭉치 배포(유상)
  - ACE 2005 말뭉치 크기
  - Training data(어휘): 영어 26만, 중국어 20.5만, 아랍어 10만
  - Evaluation data(어휘): 영어, 중국어, 아랍어 각 5만

(2) *시간 정보 주석 표준안 및 말뭉치*

① DAML-Time, OWL-Time
  - http://www.isi.edu/~hobbs/owl-time.html
  - DAML-Time을 OWL-Time으로 통합함.
  - 웹 페이지 내에 표현된 시간 정보와 웹 서비스와 관련된 시간적 특성 추출

② TIMEX, TimeML, ISO-TimeML
  - http://www.timeml.org
  - ISO의 ISO 8601, MITRE의 TIMEX3와 TIDES의 TimeML의 중심 연구자가 대부분 참여하여 통합화한 ISO-TimeML 제안중.
  - 사건(event)을 구성하는 시간개체표현(temporal entity)과 사건성 표현(eventuality) 정보와 이들 간의 링크 관계 기술

(3) *시간 정보 주석 말뭉치*

① TimeBank 1.2 : 신문기사 183개

② AQUAINT TimeML 1.0 Corpus (TimeML 1.2.1) : 신문기사 73개

③ TempEval07 : 시간개체표현(temporal entity)와 사건성표현(eventuality) 간의 링크 설정을 평가하기 위한 평가용 말뭉치

(4) 공간 정보 주석 표준안

- SpatialML 2.0
  - http://www.language-archives.org/item/oai:www.ldc.upenn.edu:LDC2008T03
  - guideline (ACE 2005 English SpatialML Annotation)을 위 사이트에서 제공하며, 배포 중인 주석 말뭉치는 아직 없음.

2.2.2.2. 국내 연구현황

○ 개체, 시간 공간 정보 분류는 경험에 의존하여 전문 분야에 따라 시소러스 형태의 분류 체계 형식을 띠고 이루어짐. 이를 토대로 기관별로 말뭉치를 구축한 실제 사례는 많으나, 분류체계·분류명 간에 상호 호환성이 결여됨.

○ 한국어를 대상으로 하는 개체명(named entity) 인식은 1990년대 중후반부터 자연언어처리 및 정보검색을 위해 연구가 진행됨.

○ 개체 정보 주석 표준안 및 말뭉치

- ACE 기반 관계 추출 방법론 연구와 소규모 적용이 연구실 단위에서 이루어짐.

○ 시간 정보 주석 표준안 및 말뭉치

- 미 Georgetown 대학에서 MITRE group에 합류한 한국 학자가 <TIMEX3>를 이용한 소규모 주석말뭉치를 구축함.
- ISO-TimeML의 작성팀을 국내 학자가 이끌고 있음.

○ 공간 정보 주석 표준안

- 미 Georgetown 대학에서 MITRE group에 합류한 한국 학자가 SpatialML 작성에 합류함.

2.2.3. 화행정보부착 말뭉치 구축

2.2.3.1. 국외 연구현황

○ 발화나 문장의 표면의미와는 별도로 화자 의도를 명시적으로 드러내는 방편으로서의 화행에 연구가 언어학, 철학, 심리학, 전산학에서 오래 연구되어 왔음.

○ 이를 전산적으로 구현해 보려는 노력의 일환으로 담화/화행 말뭉치 구축이 시도되고 있음.

○ 언어의 연구가 궁극적으로 발화자의 의도를 파악하는 데까지 이르러야 어느 정도 완결된다는 점에서 화행 연구는 필수적인 부분으로 인식되고 있음.

○ UPenn의 LDC 몇 가지 형태의 Swichboard 말뭉치가 있음.

○ 1990년대 후반들어 일군의 담화연구 학자들이 "담화 연구 이니셔티브(Discourse Resource Initiative; DRI)"를 결성하여 담화행위 다차원 주석을 위한 일반적인 방식으로 DAMSL (Dialogue Act Markup using Several Layers, Allen and Core, 1995; Core et al., 1998)를 제안하였음.

○ Switchboard DAMSL annotation project (Stolcke et al. 2000)에서는 전화대화를 대상으로 구축 방법을 정립.

○ 그 밖에도 TRAINS, VERBMOBIL, the Edinburgh Map Task Corpus, SPAAC (Leech and Weisser 2003) 등의 분석 기법 및 도구들이 개발되어 있음.

○ 유럽국가들 사이에서는 말뭉치 주석 도구를 개발하기 위한 표준으로 MATE (= Multi-level annotation, tools engineering)를 제안하였음. 언어의 여러 층위, 특히 대화행위까지도 표기가 가능한 작업용 소프트웨어임. (http://mate.nis.sdu.dk/).

○ ISO에서는 화행정보 표시에 대한 국제적 표준수립에 착수함.

○ 화행정보 부착 말뭉치 중 현재까지 구축된 대표적인 국외의 말뭉치는 다음과 같음

(1) Switchboard-1 Telephone Speech Corpus (LDC97S62)

- http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97S62
- Switchboard corpus는 1990-1년도에 최초 구축된 것으로 발표되고 그 후 수정을 거침.
- 2400개의 양방 대화로 구성되어 있으며 미국 전역에 걸친 543명의 화자(남 302, 여 241)대화.
- 로봇 교환원의 지시에 따라 주어진 주제에 대하여 대화를 하는 방식으로 대화 주제는 70여 개에서 선택
- 유료 배포 (화행 태그 부분은 무료)

2.2.3.2. 국내 연구현황

○ 일부 대학 연구실 등에서 개별적으로 구축하였거나 구축하고 있을 가능성이 있으나, 외부로 공개된 것은 없음.

○ 서강대학교에서는 호텔 예약대화와 관련한 말뭉치가 구축되어 있고, 그

와 관련된 연구가 발표되었음.

○ ETRI에서도 최근 들어 소량의 화행 말뭉치가 구축됨.
- 2006~7년 TV 가이드용 대화 말뭉치 태깅(화행,담화 등) 200세트 정도
- 2007년 텔레매틱스 환경 대화말뭉치 태깅(화행,담화 등) 200세트 정도
- 2005년 지식DB 사업을 통해 대화 말뭉치 구축 (A/S 센터, 시민상담, 호텔, 쇼핑, 티케팅, 여행, 날씨 등 각 상황별 100개 말뭉치 구축, 일부 자료에 대한 화행 태깅)

○ 연세대학교에서 세종 구어 말뭉치를 구축하는 과정에 화행 태그를 일부 시도한 것으로 알고 있으나 더 자세한 내용은 파악되지 않음.

○ 일부 회사에서 대화시스템 개발 과정에 관련 말뭉치를 구축했다고는 하나 확인되지 않음.

○ 그 밖에도 일부 대학 연구실 등에서 개별적으로 구축하였거나 구축하고 있을 가능성이 있으나, 현재로는 파악되지 않음.

2.2.4. 논조정보부착 말뭉치 구축

2.2.4.1. 국외 연구현황

○ 논조정보분석(Sentiment Analysis) 분야에서는 특정 테마에 대한 사용자의 긍정/부정 (positive/negative) 감정을 대표적으로 나타내는 어휘목록을 작성하여, 이에 기반한 텍스트의 극성(polarity)을 분석하는 방법론(bag of words method)이 초기 연구경향의 특징임.

○ 논조정보분석을 위해 긍정/부정 정보가 부착된 말뭉치가 필수적임.

○ 논조정보부착 말뭉치를 통해 각 영역(domain)별로 긍정/부정과 관련된 어휘를 (반)자동 방법을 통해 구축할 수 있음.

○ 논조정보부착 말뭉치 기반 기계학습(machine learning) 방법을 통해 문서의 극성을 분석하는 방법에 대한 연구가 진행되고 있음.

○ 가장 최근의 연구경향은 문서의 극성을 분석하는 것이 아니라, 문서를 구성하는 문장이 주관적인 내용을 담고 있는지, 객관적인 내용을 담고 있는지를 분석하여, 주관적인 문장 중 긍정/부정의 의견을 가려내는 방법에 대한 연구가 진행되고 있음.

○ 어휘 의미망도 논조분석을 위해 확장하고 있음. SentiWordNet이 이의 대표적인 예임.

○ 논조정보 부착 말뭉치 중 현재까지 구축된 대표적인 국외의 말뭉치는 다음과 같음.

(1) Blog06
- http://ir.dcs.gla.ac.uk/testcollections/access_to_data.html
- 글래스코우 대학에서 배포하는 25GB 분량의 말뭉치
- 다양한 토픽에 대한 블로그에 대해 긍정, 부정, 혼합의 태그가 붙어 있음.
- 유료배포

(2) Congressional floor-debate transcripts
- http://www.cs.cornell.edu/home/llee/data/convote.html
- 국회분야의 구어체 문장에 대한 논조태그 부착 말뭉치

(3) Cornell movie-review datasets
- http://www.cs.cornell.edu/people/pabo/movie-review-data/
- 문서단위로 보면 1천 개의 긍정적인 내용의 리뷰와 1천 개의 부정적 내용의 리뷰가 존재함.
- 문장단위로 보면 5,331개의 긍정적인 문장과 5,331개의 부정적인 문장의 데이터베이스가 존재함.
- 문장의 주관성/객관성을 파악할 수 있는 5천 개의 주관적인 문장, 5천 개의 객관적인 문장으로 구성되어 있는 데이터베이스도 구축되었음.

(4) Customer review datasets
- http://www.cs.uic.edu/~liub/FBS/CustomerReviewData.zip
- Amazon과 Cnet 사이트 등에서 다운로드한 5개의 전자제품에 대한 리뷰로 구성
- 각 문장이 주관성이 있는지 태깅하고, 만약 주관적 문장이라면 긍정/부정 등의 태그를 부착함.

(5) Economining

- http://economining.stern.nyu.edu/datasets.html
- 뉴욕대학의 Stern School에서 관리
- Amazon 사이트의 리뷰에 대한 논조태깅

(6) French sentences
- http://www.psor.ucl.ac.be/personal/yb/Resource.html
- 벨기에-프랑스어 뉴스말뭉치로부터 추출한 702개의 문장으로 구성된 소규모의 말뭉치

(7) MPQA Corpus
- http://www.cs.pitt.edu/mpqa/databaserelease/
- 535개의 뉴스 기사에 대해 문장단위 및 문장 이하의 단위까지 논조정보를 부착한 말뭉치
- 논조정보 이외에 믿음, 감정 등의 추가적인 정보까지 부착되어 있음.

(8) Multiple-aspect restaurant reviews
- http://people.csail.mit.edu/bsnyder/naacl07
- 레스토랑에 대한 4,488개의 리뷰로 구성되어 있음
- 단순히 긍정/부정 등의 논조정보만이 아니라, 음식, 서비스, 가치 등과 같은 여러 자질(feature)에 대한 긍정/부정의 정보가 부착되어 있음.

(9) Multi-Domain Sentiment Dataset
- http://www.cis.upenn.edu/~mdredze/datasets/sentiment/
- Amazon 사이트의 리뷰에 대한 논조태깅

(10) NTCIR multilingual corpus
- 일본어, 중국어, 영어 신문기사에 대한 논조태깅
- NTCIR 논조분석시스템 성능경연대회의 학습말뭉치로 사용됨.

(11) Review-search results sets
- http://www.cs.cornell.edu/home/llee/data/search-subj.html
- 야후 검색엔진에서 'review'란 단어로 검색한 결과에 대한 논조태깅

2.2.4.2. 국내 연구현황

○ 일부 검색엔진 관련 업체에서 논조분석 시스템의 개발을 위해 개별적으로 구축하고 있을 것으로 추정되나, 외부로 공개된 것은 없음.

○ 논조분석 자체가 한국어처리 관련 분야에서 최근 들어서야 비로소 주목을 받고 있는 분야이므로 기구축된 공개 말뭉치는 없음.

○ 그러나 논조자동분석에 대한 산업계의 수요는 최근 들어 급증하고 있는 상황임(2008년 3월 8일 조선일보 기사 "인터넷 댓글을 분석하라 ..." 참조).

○ 따라서 각 대학의 자연언어처리 연구실 및 관련업체에서 이에 대한 관심이 급증하고 있는 실정임.

○ 그러나 관련 연구에 필수적인 논조정보부착 말뭉치를 개별적으로 구축하기에는 시간과 비용이 많이 드는 문제가 있음.

2.2.5. 다의어 의미부착 말뭉치 구축

2.2.5.1. 국외 연구현황

○ 현재까지 국외에서 구축된 의미부착 말뭉치는 주로 어휘의미보다는 문장의 구조와 관련된 의미정보가 부착된 말뭉치임.

○ 문장의 구조와 관련된 의미정보는 술어-논항 정보라고 할 수 있음.

○ 어휘단위에서 다의어 수준으로 의미를 부착한 말뭉치는 거의 없음.

○ 다의어정보 부착 말뭉치 중 현재까지 구축된 대표적인 국외의 말뭉치는 다음과 같음.

(1) *The Hinoki Sense Bank*
- 일본어 의미부착 말뭉치
- 3백만 단어 규모, 어휘의미 부착
- 문장단위 의미정보 뿐만 아니라 어휘단위 의미정보가 부착된 외부로 공개된 최대규모의 말뭉치

(2) *DSO Sense Tagged Corpus*

- http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97T12
- 싱가폴 대학에서 구축하여 현재 LDC(Linguistic Data Consortium)에서 비회원에게 500 USD의 사용료를 받고 공개 중.
- 영어 문장에 가장 빈번히 등장하는 의미모호성이 있는 121개의 명사와 70개의 동사에 대해 의미를 태깅한 말뭉치
- Brown Corpus와 월스트리트 저널에서 추출한 약 192,800개의 문장에 대해 위 단어에 대한 의미태깅이 이루어짐
- 의미태그는 워드넷 1.5의 의미정의를 따름

(3) *Penn Prop Bank*

- http://www.cis.upenn.edu/~~ace/
- 통사정보만 부착된 Penn Treebank에 술어-논항 정보가 부착된 말뭉치
- 문장 단위의 의미정보는 부착되어 있으나, 어휘단위의 의미정보는 부착되어 있지 않음.

(4) *LinGo Redwoods Corpus*

- 통사정보가 부착된 10,000개의 문장에 대해 술어-논항 정보를 추가적으로 부착한 말뭉치
- Penn Prop Bank와 마찬가지로, 문장단위의 의미정보만 존재하고, 어휘단위의 의미정보는 부착되어 있지 않음.

2.2.5.2. 국내 연구현황

○ 세종계획사업을 통해 1천1백만 어절 규모의 의미정보 부착 말뭉치가 구축되었음.

○ 그러나 세종 의미정보부착 말뭉치는 동형이의 형태를 지니는 명사 위주(일부 동사 포함)로 구축된 말뭉치임.

○ 울산대 한국어처리연구실에서는 사전 해독용 컴퓨터를 개발하기 위한 목적으로, 표준국어대사전의 뜻풀이 전체(587,832개 뜻풀이, 약 4백만 어절)를 대상으로 명사, 동사, 형용사, 부사 등에 대해 다의어 수준의 의미를 태깅하였음.

2.3. 호환성을 갖춘 한국어 어휘의미망 구축

2.3.1. 국외 연구현황

○ 자연언어를 이해하고 생성하는 데 필요한 지식의 처리를 위해 1980년대 중반부터, 영어를 대상으로 한 WordNet(이하 PWN), 중국어를 대상으로 한 HowNet, 일본어를 대상으로 한 NTT 어휘대계 등 세분화된 어휘의미(fine-grained word sense)를 단위로 한 범용적 어휘의미망을 개발해 옴.

○ PWN 1.5가 발표된 1990년대 중반부터는 이를 참조로 한 각 언어의 어휘의미망이 개발되기 시작하면서 다국어 연계성을 확보하게 됨. 대표적인 예로는 서유럽 8개 언어를 대상으로 한 EuroWordNet(이하 EWN), 동유럽 6개 언어를 대상으로 한 BalkaNet(이하 BWN) 등을 들 수 있고, 계속 확산되면서 현재 한국어를 포함한 30개 언어를 대상으로 약 50개 어휘의미망이 개발됨. HowNet과 NTT 어휘대계도 PWN과 사상(mapping)을 시도함.

명칭	구축기관	의미/개념 vs 어휘 수	구축 품사
PWN	Princeton University	117,417s/155,297w	명,동,형,부
EWN	European Community	277,068s/484,466w	명,동
BWN	BalkaNet Consortium	78,165s/125,604w	명,동
NTT 어휘대계	NTT Cooperation	2,710n/ 40만w	명
HowNet	Chinese Academy of Sciences	95,690s/ 81,062w	명,동,형

○ 어휘의미망과는 별개로 1980년대부터 구축되어 온 SUMO, CYC, Mikro-Kosmos 등 자연언어 독립적인 상위개념망(upper ontology)과 PWN의 상위노드와의 사상도 활발히 수행됨.

명칭	구축기관	의미/개념 vs 어휘 수
Mikro-Kosmos	New Mexico State Univ.	6,000n
CYC	CYC corp.	6,000n
SUMO	IEEE SUMO Working Group	1,000n

명칭	구축(자)	분야	언어
ArchiWordNet		Architecture	Italian
WN for Aviation-domain	Davide Turcato et al.	Aviation	English
Medical WordNet	Christiane Fellbaum et al.	Consumer Health	English
CoreLex	DFKI-Language Technology	ontology and semantic database of 126 underspecified semantic types, covering around 40,000 nouns	English
LexicalFreenet	Doug Beeferman	Web-based thesaurus	English
Mimida-project	Maurice Gittens	WordNet-based mechanically-generated multilingual semantic network	20 languages
WordWeb 2	WordWeb	thesaurus/dictionary for Windows	English
OntoLing	Armando Stellato	Ontology editing tool Protégé	English

○ 이전에 존재했던 사전, 분류체계, 시소러스의 특성을 수용하였고, 2000년대 전후로 활발해진 전산학 분야의 온톨로지 구축에 중요한 모형을 제공함. 특히 전문분야 온톨로지를 개발하는 데, PWN이 초기 분류 기준 및 자질을 제공하면서 호환성의 범위를 대폭 확장함.

○ 어휘의미망은 어휘 중의성 해소(word sense disambiguation), 정보 검색(information retrieval), 다양한 수준의 의미정보 부착 말뭉치를 구축하는 데 의미 분류의 기준을 제공함.

○ 인간의 말을 이해하고, 적절히 반응하는 '똑똑한 기계'를 만들기 위해서는 인간의 언어 및 지식표상 모형의 명세화가 선행되어야 하는데, 기왕에 만들어진 어휘의미망이 중요한 역할을 할 수 있으리라고 기대하면서, 다양한 응용 분야에서 파생과 활용을 시도함. 특히 시맨틱웹(Semantic-web)으로 대변되는 의미기반 웹서비스와 지식처리와 밀접히 관련됨.

2.3.2. 국내 연구현황

○ 국내의 자연언어처리 연구가 80년대 중반에 시작한 만큼 지식표상체계의 구축도 90년대 중·후반에 출발함. 자연언어처리 시스템에 필요한 의미분석을 해야 한다는 실질적인 목적을 가진 전산학자들에 의해 주도함. 주로 명사를 대상으로 작은 크기의 어휘의미망이 시험적으로 구축되기 시작함. PWN, EWN을 참조 모델로 함.

명칭	중심구축 기관	중심구축자 전공	구축기간	구축방식/ 참조모델	의미/개념(n) vs 어의(w) 수	구축 품사
한국어 명사워드넷	호남대학교	전산학	1994-1995	직접	20,000w	명
한국어 시소러스	포항공대	전산학	1997-2000	참조/PWN	18,362n vs. 21,390w	명
다국어 어휘 데이터베이스	고려대학교	언어학	2000-2006	참조/EWN	5,500w	명

○ 어휘의미망의 형태를 띠지는 않으나, 관련된 언어자원으로는 국립국어원에서 편찬한 『표준국어대사전』(이하, 표준)과 국어정보화 사업 1단계 결과인 『세종전자사전』(이하, 세종)이 구축됨.

명칭	중심구축 기관	중심구축자 전공	구축기간	의미/개념(n) vs 어의(w) 수	구축 품사
표준국어대사전	국립국어원	언어학	***	*****20,000w	모든 품사
세종 전자사전	서울대학교	언어학	1998-2007	581n vs. 540,000w	모든 품사

○ 2008년 11월 현재 3개의 중대형 크기의 범용 어휘의미망이 구축됨. 전산학과 언어학이 공동으로 개발에 참여하였으며, 국어사전의 정의문을 직간접적으로 연계함. NTT어휘대계, PWN을 참조 모델로 삼거나, 국어사전의 정의문을 이용한 직접 구축 방식을 이용해 개발함.

명칭	중심구축 기관	중심구축자 전공	구축기간	구축방식/참조모델	의미/개념(n) vs 어의(w) 수	구축품사	참조 사전
CoreNet	KAIST	전산학/언어학	1995-2004	참조/NTT어휘대계	2,938n vs. 62,632w	명,동,형	우리말 큰사전
U-Win	울산대학교	전산학/언어학	2002-2007	직접, 참조	46,339n vs. 약250,000w	모든 품사	표준
KorLex 1.5	부산대학교	전산학/언어학	2004-현재	참조/PWN	130,639n vs. 147,906w	명,동,형,부,분류사	표준

○ CoreNet은 국내 최초로 만들어진 어휘의미망으로, 상위 온톨로지와 단말 어휘의미와의 구분이 잘되어 있다는 장점이 있으나, 어휘의미망의 크기가 작고, 다국어 연계성이 떨어짐.

○ U-Win은 어휘의미망의 크기가 가장 크고 한국어 어휘의미구조를 가장 충실하게 반영하고 있으나, 다국어 연계성이 없음.

○ KorLex는 범용 어휘의미망으로서 사용하기에 크기가 충분하고, PWN을 참조모델로 하여 다국어 연계성이 뛰어나지만, 영어에 경도됨.

○ 국가에서 주도적으로 개발한 언어자원인 표준, 세종 사전과 기존 어휘의미망과의 통합적인 연계가 필요함. 부분적으로 U-WIN과 표준을 통합한 어휘의미망을 개발하고 있으나, 크기가 작고 U-WIN의 단점인 다국어 연계성이 확보되지 못함.

3. 전문용어 및 언어자원 표준화 분야

3.1 전문지식의 보편화와 자생적 융합을 위한 전문용어 표준화 체계 구축

○ 전문용어는 학술용어, 번역의 대상 용어, 일상용어, 물류용어 등으로 그 쓰임에 따라 구분될 수 있음.

○ 전문용어를 일상용어화하기 쉬울 때, 그 국가와 사회가 좀더 과학기술이 일상 생활화할 수 있음. 학술용어가 일상용어와 동떨어진 언어일 때, 일상용어를 사용하는 국민 다수의 사고가 선진화하는 속도는 떨어진다고 봄.

○ 전문용어는 한 개의 개념을 여러 나라 언어로 번역하여야 한다는 점에서 실용성을 가짐.

○ 전문용어는 인터넷 검색의 색인 혹은 키워드로 쓰임. 전문용어의 표준화가 되지 않으면 한 개의 개념에 대하여 여러 형태의 단어를 쓰게 되어, 검색이 충분히 이루어지지 않으며, 관련된 지식과 정보가 통합되지 않음.

○ 21세기의 IT 기반 지식정보화 사회에서는 전문용어가 지식의 표준화를 위한 국가적 산업인프라로 인식되고 있음. 지식정보 유통의 다원화로 인해 외국어 특히 영어 용어에 대한 수용 방식에 있어서 고유어, 한자어, 원어 차용 등 용어사용의 개인화로 인해 다양한 용어 사용 방식이 난립하고 있음. 이로 인해 현재 각 세계 선진국들은 자국의 전문용어의 정비는 물론 새로 생성되는 전문용어를 효율적으로 관리할 수 있는 전문용어관리시스템의 확립과 개발 공정의 체제 구축에 전력을 기울이고 있음.

3.1.1 국외 현황

○ 국제전문용어조직으로 InfoTerm과 Termnet이 있음. InfoTerm은 각 나라 대표가 참여하는 전문용어의 UN과 같은 조직이며 TermNet은 모든 전문용어에 관여되는 조직들의 협력체임.

○ 국제표준화기구 (ISO)에는 1938년도부터 ISO/TC37 전문용어표준화 위원회를 운영하여 왔음. 최근 ISO/TC37은 “전문용어, 언어 및 컨텐츠 자원” 표준화 위원회로 개칭하였음.

○ 지역별 표준화 상황은 다음과 같음:

(1) 유럽: EAFT (European Assocation Forum of Terminology) 등 전문용어 관련 단체가 집중적으로 활동하고 있으며, 유럽 통합의 한 축을 이루고 있음.

(2) 일본:

- 전문용어의 자국어화로 세계적 기술수준에 도달
- 정부와 학회의 협력 사업을 통해 전문용어 표준화 대역 목록인 “학술용어집”을 32개 분야에 걸쳐 완성
- 전문용어의 전통적 연구집단이 형성
- 대학에서까지 교과서에서 “학술용어집”의 전문용어를 채택
- 다국어 대역 전자사전의 구축을 통한 일본어의 위상 정립
- 번역용 사전의 세계 첨단을 구가

(3) 중국: 국가 지식정보 관리 차원에서 전문용어 표준화 및 국제표준 규격을 적극적을 도입함.

3.1.2 국내 현황

○ 21세기 세종계획에 의하여 “전문용어 정비” 사업이 이루어짐 (1998-2006). 그 결과 기초적 과학기술분야 학술용어 10만 여 용어에 대한 데이터베이스가 이루어짐. 기초과학분야인 수학, 물리, 화학, 생물 등을 먼저 정비하여 응용과학기술분야가 이를 기반으로 정비를 하도록 하는 계층적 접근을 하였음.

○ 전문용어 정비사업은 KAIST 전문용어언어공학연구센터 (Korterm)와 국내의 과학기술 분야 학술단체와의 계약에 의하여 진행이 됨. Korterm은 국내외 용어를 수집하고 이를 과학기술분야 학술단체에 제공을 하고, 용어의 정비 지침을 마련하여 학술단체의 용어편집진이 일관성 있는 용어를 정비하도록 유도함. 이의 국어학적 분석은 연세대 국어국문학과에서 담당함.

○ 한국학술단체연합회는 한국학술진흥재단의 지원으로 일정 요건을 갖추면 인문, 사회분야를 비롯한 모든 학술단체의 학술전문용어 정비 및 표준화를 지원함(2004 - 2007).

○ 남북한 전문용어와 관련하여 산자부와 표준협회의 지원으로 ISO 2382 전문

용어의 남북한 표준안이 마련됨(2006).

○ 번역 및 통역대학원에서 용어 연구 및 개발이 이루어지고 있음.

○ 전자상거래, 물류, 웹서비스를 위한 온톨로지, 번역 등의 필요와 목적으로 민간 표준이 개발되고 있음.

○ KS 표준으로는 각 국제표준 분과위원회의 용어표준을 중심으로 전문용어 표준화가 이루어지고 있으나, 국립국어원, 한국학술진흥재단, 기술표준원, 특허청 등의 관련 기관 간의 통합된 협력체계 구축이 이루어지고 있지 않음.

(1) 한국학술단체총연합회

○ 구축시기
- 1 차 : 2003 년~2005 년
- 2 차 : 2006 년~2007 년

○ 구축목적 : 학술 전문용어 정보 및 표준화 사업

○ 구축규모 : 366,591개

○ 구축분야 : 인문과학, 물성과학, 생명과학, 에체능 4개 분야 39개 세부분야 40개 학회

○ 문제점/특징
- 전문용어의 체계적인 생성 및 관리 환경 부족
- 용어의 세부 분류별 정보가 통제되지 않고 있음.

번호	분 야	용어수	번호	분 야	용어수
1	기호학	1,856	21	대기과학	1,034
2	서양사학	4,991	22	해양학	10,164
3	동양사학	4,049	23	농림학	25,612
4	종교학	3,910	24	수산학	10,166
5	고고학	3,528	25	과학기술학	5,096
6	지리학	9,363	26	조경학	9,544
7	교육학	19,946	27	의학	35,312
8	만화애니메이션학	4,468	28	치의학	18,274
9	음악학	1,245	29	약학	3,072
10	연극/영화학	4,000/1,905	30	무용학	10,102
11	정치외교학	4,797	31	체육학	8,332
12	사회학	2,745	32	기계공학	19,512
13	법학	1,894	33	산업공학	4,977
14	행정학	10,203	34	콘크리트공학	(5,218)
15	신문방송학	4,978	35	자원공학	8,436
16	경제학	6,655	36	건축공학	2,000
17	무역학	5,322	37	컴퓨터공학	5,668
18	경영학	9,605	38	항공우주공학	8,922
19	지질학	9,243	39	인지과학	51,570
20	천문학	8,877	TOTAL	39 분야(40 학회)	366,591

(2) 국립국어원

○ 구축시기 : 1998년~2007년

○ 구축목적

- 21 세기 세종계획 사업의 일환
- 전문용어의 수집/ 정비/ 관리/ 활용의 체계화
- KAISTI 전문용어언어공학연구센터(KORTERM)에서 관리

○ 구축규모 : 100,000 여개

○ 구축분야 : 인문사회(경제), 과학기술

○ 문제점/특징

- 전체 분야에 대한 지속적 연구와 관리가 이루어지지 못함.

- 기존 출판도서의 활용이 많음.

(3) 정보통신기술협회

○ 구축시기 : 1995 년~

○ 구축목적 : 3 개부처(정보통신부/ 문화관광부/ 기술표준원) 를 중심으로 한 정보통신용어 표준화 작업

○ 구축규모 : 23,000 여 개

○ 구축분야 : 전기통신, 무선/ 방송, 정보기술, 데이터통신 및 S/W

○ 문제점/특징

- 분야 정보의 불분명
- 소량의 표준화 전문용어
- 국어순화용어 사용 미비

(4) ETRI(한국전자통신연구원)

○ 구축시기 : 2004 년~

○ 구축목적 : 한국어 표준형 음성 DB, 다국어(중국, 일본, 영어 등) 기계번역DB 구축

○ 구축규모 : 1,500,000여 엔트리(국내 최대 규모 보유)

○ 구축분야 : 특허 문서를 중심으로 과학 전분야

○ 문제점/특징

- 특허 정보만을 대상으로 추출
- 학술단체등과의 연계 부족
- 용어의 분류 정보가 미흡
- 웹상에서 분야별로 서비스 중(판매)

분류	사전명	구축 현황 (엔트리 수)	서비스 되는 형태	기타 특징
일반 대역 사전	영한 번역 사전	105,000	• 사전만 별도로 공개 서비스 되고 있지 않음 • ETRI 영한 자동번역 엔진에 연동	• 영어 표제어의 한국어 대역어 각각에 WordNet의 Synset이 어휘의미코드로 부여되어 있음
	한영 번역 사전	150,000	• 사전만 별도로 공개 서비스 되고 있지 않음 • ETRI 한영 자동번역 엔진에 연동	• 2005, 2006년 ICU, KAIST 에서 기술이전 함
	한중 번역 사전	150,000	• 사전만 별도로 공개 서비스 되고 있지 않음 • 한중 자동번역 엔진에 연동	• 2004년 ㈜코난테크놀로지 에서 기술이전 함
	중한 번역 사전	220,000	• 사전만 별도로 공개 서비스 되고 있지 않음 • 지식 DB 사업 결과물로• 유료 배포 중 • 중한 자동번역 엔진에 연동	• 지식 DB 과제의 결과물로써, DB 배포 중 (http://slrdb.etri.re.kr/db/guide_01.asp?leftNum=1) • 2007년 ㈜LNI소프트에 중한 자동번역 엔진 목적코드를 기술이전 함
	일한 번역 사전	100,000	• 사전만 별도로 공개 서비스 되고 있지 않음 • 일한 자동번역 엔진에 연동	• 2001년 클릭큐, LNI에 일한/한일 양방향 자동번역 엔진 기술이전 함
	한일 번역 사전	100,000	• 사전만 별도로 공개 서비스 되고 있지 않음 • 한일 자동번역 엔진에 연동	• 2001년 클릭큐, LNI에 일한/한일 양방향 자동번역 엔진 기술이전 함
전문 용어 사전	영한 전문용어 사전	1,650,000	• 사전만 별도로 공개 서비스 되고 있지 않음 • 지식 DB 사업 결과물로써, 유료 배포 중 • 영한 특허문서/기술논문 자동번역 엔진에 연동	• 한국어 대역어 포함, 전자, 기계, 화학, 컴퓨터, 의료 전문용어로 구성되나, 분야별로 구분은 없음. • 지식 DB 과제의 결과물로써, DB 배포 중 (http://slrdb.etri.re.kr/db/guide_01.asp?leftNum=1)
	한영 전문용어 사전	2,000,000	• 사전만 별도로 공개 서비스 되고 있지 않음 • 지식 DB 사업 결과물로써, 유료 배포 중 • 영한 특허문서/기술논문 자동번역 엔진에 연동	• 영어 대역어 포함, 분야 구분은 따로 없음 • 지식 DB 과제의 결과물로써, DB 배포 중 (http://slrdb.etri.re.kr/db/guide_01.asp?leftNum=1)

분류	사전명	구축 현황 (엔트리수)	서비스 되는 형태	기타 특징
시소러스	시소러스	80,000	• 별도로 공개 서비스되고 있지 않음	• 명사 및 용언 의미 개념망 포함 • ETRI 지식 마이닝팀 보유
	한국어 의미 사전	360,000	• ETRI 한영/한중 자동번역 엔진에 연동	• 명사 의미 정보 포함
기타	한국어 형태소 사전	160,000	• 별도로 서비스되고 있지 않음 • ETRI 형태소 분석기 사전으로 사용	• 형태소 분석용 부가지식 포함

(5) 한국과학기술정보연구원(KISTI)

○ 구축시기 : 2003 년

○ 구축목적 : 과학기술분야 기반 언어자원 DB 정비와 관리 체계 통합화

○ 구축규모 : 320,213 개

○ 구축분야 : 과학기술 분야

○ 문제점

- 기존의 전문용어 수집 및 스키마 통일
- 출판도서를 이용한 전문용어 수집 및 입력

3.2 국어 언어자원관리 표준 구축 및 생활 전문용어 구축

언어자원표준은 전통적으로 전문용어의 표준에서 시작되어, 일반적 언어의 부가가치화와 형식 표준을 의미한다. 모든 컨텐츠에 내재한 언어의 형식 표준을 추구하고 있다.

3.2.1 국외 현황

○ ISO/TC37은 전문용어와 언어자원, 그리고 컨텐츠 자원의 표준화를 목표로

하고 있다. 최근 DCR (Data Category Registry) 표준을 확립하여 언어자원에 대한 명칭까지 국제표준화하려고 하고 있음. 주요 표준화 분과 및 과제는 다음과 같음.

(1) SC1: 전문용어 표준화 제정에 대한 원칙

(2) SC2: 사전을 만들 때 필요한 형식, 기호 등 메타데이터와 일반사전 및 전문용어 사전 작성 시에 필요한 지침 및 인증 절차: 출판사, e-Learning 관련 회사 및 단체가 참여하고 있음.

(3) SC3: 전문용어의 컴퓨터 활용 표준

(4) SC4: 언어자원 작성 형식 지침

- 언어자원의 디지털화를 위한 기본 형식으로서 LAF (Linguistic Annotation Framework)를 설정한다. W3C의 XML 관련 표준을 활용한다. 예를 들어, 텍스트 안에서 각 단어나 문장의 연결 참조를 표현하기 위하여 XML에서 정한 표준을 적용하여, 언어자원의 추상구조를 실현하며 또한 기계가 읽을 수 있도록 함.
- 언어계층의 형식화: 형태, 통사, 의미, 대화구조, 시제, 공간에 대한 마크업 표준을 정함.
- 다국어화를 위한 구조: 번역을 위한 문서 구조를 정의함
- 자연언어처리용 사전 구조를 정의함.

○ 최근 유럽에서는 EU 공동프로젝트로서 CLARIN (EU FP과제)과 FlareNet를 추진 중이다:

- CLARIN: A European Network for building and strengthening collaborative infrastructures for scientific research where Language Resources and Language Technologies (LRT) are relevant (언어자원과 언어기술이 관련된 과학적 연구를 지원하기 위한 협력 인프라 구축과 강화를 위한 유럽 네트워크)
- FlareNet: Fostering Language Resources Network (언어자원의 개발 및 사용 주체 네트워크를 확립하고, 언어자원 평가체계 구축)

○ 아시아 자연언어처리협회 (AFNLP)를 중심으로 언어자원위원회가 구성되어, 언어자원 상호운용성 및 교환을 추진하고 있음.

○ 일본 NICT (National Institute of Information and Communication Technology)에서는 Language Grid라는 개념으로 언어자원을 응용 시스템에 묶어 쓸 수 있도록 추진하고 있음.

3.2.2 국내 현황

○ 국어정보베이스 (KAIST 1995-1998) 과제가 SERI 국어공학센터 주관으로 이루어져, 그 당시 한국어 품사표준과 말뭉치 형식표준을 한국 최초로 정한 바 있음.

○ 21세기 세종계획 (1998-2007)에서 이와 비슷한 노력을 하였음.

○ 국가지정 언어자원은행 (BORA 2003-2007)이 과학재단 재원으로 이루어져 언어자원을 표준화하여 국내외적으로 배포하는 노력을 하였음.

○ ISO/TC37 한국 전문위원회를 중심으로 국제화에 대응하여 왔음.

3.3. 언어처리 기술 평가를 위한 표준화된 평가 세트 구축 현황

3.3.1. 국외 현황

○ 국외의 경우 정보검색 기술의 발전을 위하여 1992 년부터 매년 개최되는 가장 주요한 학술대회 중 하나인 TREC(Text REtrieval Conference)을 열고 있음. (<표 9> 참조)

○ TREC은 정보검색의 여러 주요 기술에 대하여 집중적으로 연구하기 위해 분야마다 트랙(track)을 두어 진행하여 왔음.

○ 새로운 기술에 대하여 새로운 트랙을 개설하고 충분히 연구된 주제에 대한 트랙은 중단하는 방식으로 운영되어 왔으며, 그 결과 정보검색의 많은 주요한 문제에 대하여 깊이 있는 실험을 수행하였고, TREC은 정보검색 기술의 발전에 매우 큰 기여를 하였다고 인정되고 있음.

○ 정보분류와 관련하여 매일 수많은 문서가 생성되는 현대 사회에서 모든 문서를 다 검토하는 것은 매우 어려움. 따라서 문서들을 미리 기계가 여러 분야로 분류하여 놓는다면 사람은 자신이 관심이 있는 분야(범주 category)로 분류된 문서만을 읽어 볼 수 있게 되고 이는 많은 시간과 노력을 절감할 수 있도록 할 수 있기 때문에 문서 분류 기술은 최근 정보검색 분야에서 중요시됨.

○ 이러한 문서분류 시스템을 개발하기 위해서는 교사학습(supervised learning) 기반 기계학습 기법을 사용하는 것이 주된 추세임. 이를 위해

서는 상당량의 문서에 범주 레이블을 미리 붙여 놓은 평가 세트가 있어야 함.

○ 평가 세트의 구축은 많은 시간과 노력이 들어가는 작업으로서 개개의 연구 집단이 수행하기에는 너무 부담이 크며, 이러한 이유로 선진국에서는 국가기관이나 대형 연구기관에서 이를 구축하여 연구자들로 하여금 이용할 수 있게 하고 있음.

3.3.1.1. TREC

○ 미국 NIST(National Institute of Science Technology)의 관련 기술자들은 정보검색 기술의 중요성을 인식하고 이의 발전을 위하여 1992년부터 TREC 이라는 독특한 형태의 학술 진흥 모임을 기획하게 되었음.

○ TREC은 아래와 같은 측면에서 중요함.

- 정보검색 시스템의 개발을 위해서는 대용량의 평가 세트가 필요한데 이는 일반 개인 연구자들이 준비하기 어려움.
- 여러 연구 집단에서 개발한 시스템의 성능 평가를 통하여 우수한 기술을 파악하고 이를 연구 집단이나 산업계에 전파할 수 있음.
- 평가 및 비교가 신뢰성이 있고 객관성이 있기 위해서는 표준화된 동일한 평가 세트를 이용한 실험이 필요하며 TREC은 이를 위한 환경을 제공함.

○ TREC 초기 몇 년은 정보검색의 가장 기본적인 문제인 질의에 대한 적합한 문서를 검색하는 작업에 집중하였으나 90년대 말부터는 여러 가지 다양한 문제들을 다루게 되었고 이를 각 트랙별로 나누어서 진행하도록 하였음.

○ TREC은 전 세계에서 정보검색과 관련된 연구를 이끄는 대부분의 연구 집단들이 참여하는 주요한 학술 모임으로 성장하였음.

○ 이를 모방하여 일본에서는 영어는 물론 동양어로 된 문서에 대한 정보검색 기술의 활성화를 도모하기 위한 NTCIR(NACSIS Test Collection for IR Systems) 학술대회를 개최하고 있음.

<표 9> TREC 추진 현황

Tracks	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Ad-hoc	■	■	■	■	■	■	■	■	■							
Filtering				■	■	■	■	■	■	■	■					
Novelty											■	■	■			
High precision						■	■									
Spoken document						■	■	■	■							
Cross language (2003이후 CLEF로 분리)							■	■	■	■	■	■	■	■	■	■
Interactive						■	■	■	■	■	■	■				
Web								■	■	■	■	■	■			
Hard												■	■	■		
Robust retrieval													■	■		
Tera byte													■	■	■	
Video (2003이후 TRECVID)										■	■	■	■	■	■	■
QA								■	■	■	■	■	■	■	■	■
CiQA														■	■	■
Genomics												■	■	■	■	■
SPAM														■	■	■
Legal															■	■
Blog															■	■
Enterprise															■	■
Million Query																■

3.3.1.2. 해외 구축 주요 정보검색 평가 세트

○ NTCIR (NACSIS Test Collection for IR Systems): 일본의 NTCIR (NACSIS Test Collection for IR Systems) 평가 세트(Kando et al, 1999)는 NACSIS (National Center for Science Information Systems)가 주관이 되어 개발되고 있음.

- 339,483개의 문서와 100개의 질의를 포함하고 있음. 문서집합은 여러 분야의 학회논문으로 된 일본어 문서, 영어 문서, 일-영 대응문서 등으로 구성되어 있음.

- 일반문서검색, 일-영 교차언어 검색, 그리고 전문용어 인식에 대한 평가집합을 포함하고 있음.

○ BMIR-J1&J2: 일본의 NTT Data Corporation에서 BMIR-J1과 BMIR-J2를 개발하였으며, BMIR-J1은 600건의 문서와 60개의 질의어로 구성되어 있으며, BMIR-J2는 경제학 및 공학 분야의 5080건의 신문기사와 60개의 질의어로 구성되어 있음.

○ CLEF: CLEF(Cross-Language Evaluation Forum)-2000 (CLEF, 2000)은 유럽 언어에서 정보검색 시스템 평가를 위한 것으로, 유럽 언어들에 대한 다국어 검색, 교차언어 검색, 문서검색을 평가하기 위한 평가 세트를 구축하였음.

- 문서는 같은 시기의 영어, 독일어, 불어, 이탈리아어 다국어 문서집합으로, 독일어 152,694개 문서, 불어 44,013개 문서, 이탈리아어 58,051개 문서와 영어 113,005개의 문서를 포함 총 367,763개로 이루어져 있음.
- 질의는 25개로 네덜란드어, 영어, 불어, 독일어, 이탈리아어, 스페인어, 스웨덴어, 핀란드어로 만들어짐.

○ **OHSUMED:** OHSUMED 평가 세트 (Hersh, 1994)은 문서 348,566개(270개 의학 저널 1987~1991년)와 질의 106개로 구성되어 있고, 문서에는 의학분야의 분류정보가 포함되어 있음.

- 문서분류에 대한 평가 세트로 활용되고 있고, TREC-9(2000년)에서는 정보 필터링의 평가 세트로 이용되었음.

3.3.1.3. 문서분류 평가 세트

○ Reuters 평가 세트: Reuters-21578이라 불리는 이 평가 세트는 1987년의 Reuters 통신사의 뉴스를 수집하여 이들을 사람들이 수동으로 태깅을 함으로써 탄생하였음.

- 문서는 SGML 포맷으로 되어 있고 파일당 1,000 개씩의 문서를 담은 총 22개의 파일에 저장되어 있음.

○ 20-Newsgroup 평가 세트: 이 데이트 세트는 Lang에 의하여 개발된 것으로서 Usenet 뉴스 그룹에 email 메시지를 모은 것임.

- 데이터를 수집한 뉴스 그룹의 수는 20개이며 그룹마다 모든 메시지의 수는 각각 1,000개씩으로 전체 데이터 집합의 문서 수는 20,000개임.

- 각 뉴스그룹 이름이 범주의 이름이 됨.

○ RCV 평가 세트: 최근에 Reuters 사에서는 RCV라고 불리는 총 문서 수 800,000개의 거대한 문서분류 평가 세트를 구축하였음.

3.3.2. 국내 현황

○ 지금까지 국내외에서 구축된 주요 언어자원은 크게 말뭉치와 정보 시스템을 테스트하기 위한 평가 세트로 구분됨.

○ 국내의 경우 말뭉치 구축에 집중한 나머지, 정작 언어자원의 테스트와 이를 통한 정보 체제의 활성화에 대한 투자는 상대적으로 미비하였음.

3.3.2.1. 정보검색 평가 세트

○ 1998년도에 비교적 중규모의 HANTEC(한텍) 2.0이라는 평가 세트가 개발 되었음. 이는 보통 중규모 이상에서 적합성 정보를 수동으로 구축할 수 없기 때문에 일반적으로 사용되는 풀링(pooling)방법을 사용하여 적합성 정보를 구축하였음.

○ 그러나 그 이후에 평가 세트 구축이 중단되었으며 현존하는 많은 정보검색 시스템의 질적 평가를 하기에는 활성화되지 않은 측면이 있음.

- **KT-SET 1.0**
  KT-SET 1.0은 1994년에 한국통신에서 구축한 평가 세트로, 30개의 단순질의와 1,053개의 학회논문 초록만으로 이루어져 있음.
- **KT-SET 2.0**
  1996년 KT-SET 1.0을 4,414건의 문서와 50개의 질의어로 확장한 평가 세트로, 전기/전자, 컴퓨터 분야의 논문, 신문기사 등으로 이루어져 있음.
- **KRIST 평가 세트**
  1995년에 13,315건의 과기처 연구보고서와 30개의 질의어로 구성되어진 평가 세트로, 생명과학, 의용전자공학, 기계공학 등을 대상으로 구축되었음.
- **ETRI-Kemong-set**

1997년 계몽사 백과사전으로 하나의 표제어에 해당하는 것을 하나의 문서로 만든 것으로 23,113개의 문서와 46개의 질의로 구성되었으며, 각 문서는 12개의 대분류와 76개의 소분류에 대한 분류정보를 포함하고 있어서 문서분류의 평가 세트로 이용될 수 있음.

- **HANTEC 1.0**
  평가 세트 구성에 가장 기본인 문서 집합과, 질의어 집합, 각 질의어에 적합한 문서 리스트로 구성되었으며, 사회과학, 과학기술분야에 속하는 120,000건의, 짧게는 수십 바이트에서 길게는 수십만 바이트까지 다양한 크기의 문서들로 이루어져 있음.

- **HANTEC 2.0**
  HANTEC 2.0에서는 HANTEC 1.0에서 사용하던 문서들을 그대로 사용하였으며, 질의어 형식은 1.0의 형식을 유지 하였으나, 질의어의 개수를 조정, 과학기술 분야 질의어 20개를 추가하여 총 과학기술분야 30개 질의어로 구성, 전체 질의어를 총 50개로 구성하였음.

3.3.2.2. 문서분류 평가 세트

○ 한국어로 된 문서분류 평가 세트는 아직 공식적으로 발표된 것이 없어, 한국어 문서 분류 시스템의 개발 및 비교에 큰 어려움이 있음.

○ 현재 국내에는 KRTC-2003이라 불리는 단일범주분류를 위한 비공식적인 평가 세트가 있음.

○ 다른 문서분류 평가 세트와 크게 다른 점은 계층화된 범주 레이블을 이용한다는 점이며, 기존 연구에서는 거의 다루어 본 적이 없는 분류체계임.

3.3.2.3. 기계번역시스템 성능 평가를 위한 평가세트

○ 상용 영한 기계번역시스템의 번역 품질에 대한 객관적인 평가와 개발 보조를 위해 활용 가능한 평가세트 형식 설계

○ 시범 DB 구축

- 전문용어언어공학연구센터(KORTERM, 20011)
- 대상시스템: 영한 기계번역시스템

- 구축량: 총15,000여 문장

○ 평가 모델링

- 객관적 품질 평가를 위한 평가 방법론 정립
- 구축된 시범 평가세트를 이용하여 제품의 번역 결과에서 문제가 있는 언어현상 분석

○ 평가 모델링

3.3.2.4. 형태소분석을 위한 평가세트

○ 형태소분석기의 입출력 표현 양식과 정보의 종류에 대한 표준안의 필요성을 인식하여, ETRI 주관으로 1999년 제1회 형태소분석기 및 품사태거 평가대회가 개최되었음(MATEX99)

III. 전체 사업 목표

○ **21세기 세종계획**은 "우리말과 우리글을 바탕으로 하는 정보사회 건설"을 위해 세계수준의 국어 기초 언어 자료베이스 구축을 통한 우리말 정보화, 표준화된 전자사전 구축을 통한 우리말 체계화, 한민족 언어 정보화를 통한 우리말 세계화를 사업 목적으로 하여 수행되었다.

○ 국어정보화 2단계 사업은 진보된 인터넷 환경에서 21세기 세종계획 성과물을 원활하게 확산시키며, 지식사회에 걸맞는 새로운 한국어 정보 체계를 확립하고 도래하는 새로운 정보 구조에 맞는 한국어 지식 체제의 확립을 목표로 한다.

○ 이러한 목표 달성을 위해, 향후 5년간 우선 진행될 2단계 국어정보화 사업의 중점 추진 사업으로 다음의 세 분야를 선정하였다.

(1) 국어자원을 활용함으로써 **지식사회를 선도하는 한국어**
- 한국어 정보처리 인프라 구축을 위한 실용 언어 자원 구축
- 언어 산업과 연계한 언어 인프라 구축
- 국어 자원의 IT 활용을 위한 공유체제 구축

(2) 다양한 의미분석 자원을 구축하여 **미래를 준비하는 한국어**
- 다양한 의미분석 자원 말뭉치 구축
- 호환성을 갖춘 한국어 어휘의미망 구축
- 의미분석 자원 간의 연계/연동 시스템 구축

(3) 언어자원을 표준화하고 전문용어를 구축하는 **세계와 소통하는 한국어**
- 전문지식의 보편화와 자생적 융합을 위한 전문용어 표준화 체계 구축
- 국어 언어자원관리 표준 구축 및 생활 전문용어 구축
- 언어처리 기술 평가를 위한 표준화된 평가 세트 구축

○ 위 세 분야의 사업을 통한 국어정보화 2단계 사업의 비전은 다음 <그림 7>과 같다.

<그림 7> 국어정보화 2단계 사업 비전

○ **한국어 IT HUB**는 국어정보화 사업을 통해 구축된 여러 언어자원 및 언어처리 지원 프로그램을 사용자의 요구에 즉시 대응할 수 있는 **통합형 국어정보베이스**로 관리하며, 차세대 웹 기반으로 여러 계층의 자발적인 참여 및 협업이 가능한 대국민 서비스를 위한 포털 역할을 담당한다.

○ 또한, 이러한 **한국어 IT HUB** 운영에 기업이 참여할 수 있게 함으로써, 통합형 국어정보베이스를 다양한 방법으로 실용화하고 산업화할 수 있게 한다.

IV. 사업 내용

1. 지식사회를 선도하는 한국어

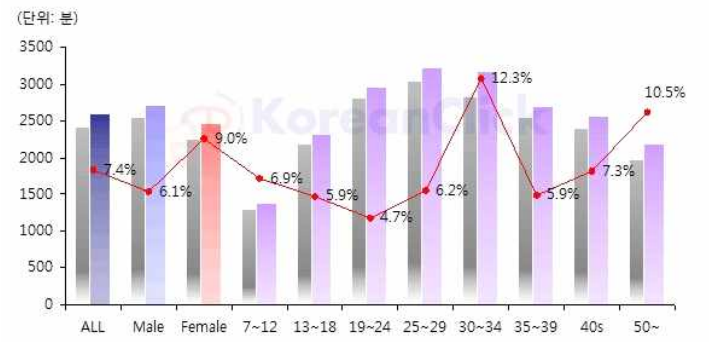
1.1. 사업의 필요성

1.1.1. 새로운 정보 수집 패러다임의 필요성

○ 오늘날 인터넷이 대중화되면서 전국민의 인터넷 노출 시간이 많아지고 있음.

○ 한국인터넷진흥원 자료에 따르면 전체 인구의 약 80%가 인터넷을 사용하고 있으며, 또 코리안 클릭의 2008년 인터넷 서비스 동향자료를 참조하면 인터넷 사용자의 한달간 평균 인터넷 이용시간은 2,000분 이상이고 월 33시간 이상 인터넷을 이용하는 것으로 알려짐.

○ 국어 등 언어에 대한 접점이 점차 책 등을 이용한 형태에서 인터넷을 활용한 온라인 형태로 바뀌어 가고 있으며 국어자원을 IT에서 활용하는 것이 중요한 이슈로 대두되고 있음.



<그림 8> 2008년 1월 연령별 인터넷 이용시간 (코리안클릭)

○ 국어정보도 일부 전문가들에 의해서만 구축되는 기존구축방식을 이용할 경우 급속도로 증가하고 변화하는 언어현상을 제대로 반영하지 못하여

계속 시대에 뒤처지는 정보가 됨.

○ 학술적으로나 실용적인 측면에서 바라보았을 때 국어정보를 통합적으로 관리하는 개념이 각기 다르고 지향하는 바가 다름을 알 수 있음.

○ 이러한 문제점을 해결하기 위해서는 실시간으로 정보가 등록되고, 이에 대한 분석이 함께 이루어져야 하는데, 개방형 IT 허브를 구축하면 이러한 문제를 상당 부분 해결할 수 있음.

○ 오늘날 한국어 정보는 과거와는 다른 양상으로 발전하고 있는데 뉴스 동영상과 동기화되어 있는 기사 전사 말뭉치나 외국어 동영상에 동기화되어 있는 SMIL로 구성된 자막정보 등은 지금까지 세종계획에서 구축된 언어자원과는 전혀 다른 형태의 모습을 띠고 있음.

○ 이러한 정보들을 다시 말뭉치로 구축하고 용례를 찾으려면 오랜 계획을 수립하고, 다시 이를 큰 비용을 들여 구축해야 하나 이 경우에는 여러 문제점이 존재함.

- 이미 구축된 자료가 있을 경우 중복되어 구축될 확률이 높음.
- 구축된 이후에는 현재 발생하는 언어 현상을 즉시 나타낼 수 없음.
- 전문용어와 같은 신조어 영역은 이미 정식 사업을 벌여 구축한 뒤에는 시대에 뒤떨어진 용어가 됨.

1.1.2. 온라인 기반 통합 국어 정보 유통 허브의 필요성

○ 구축된 정보가 유통되지 않는다면, 그 정보는 단순한 하나의 "모음"에 불과하며, 현재는 구축된 국어정보가 제대로 유통되어 활용되지 못하고 있음.

○ 따라서 구축된 모든 정보 라이브러리가 유기적으로 연동되고, 활용될 수 있는 복합적인 플랫폼이 제공될 필요가 있음.

○ 특히 최근 경향인 공유와 개방의 개념을 도입하여 온라인 기반 통합 국어 정보 유통 허브를 구축한다면 아래와 같은 긍정적인 효과를 얻을 수 있음.

○ 국어자원의 활용/수집 효율성 증대

- **집단지성을 이용한 양질의 국어자원 구축:** 지금까지 큰 비용을 들여 구축했던 한국어 정보 라이브러리, 말뭉치 정보를 집단지성에 의거하

여 집합적 지식의 개념으로 구축할 수 있게 한다면 현재 언어현상을 가장 빠르게 수집하고 이를 분석할 수 있는 기반을 마련할 수 있을 것임.

- **한국어 자원 활용을 위한 기술 인력 양성:** 한국어 정보처리는 현재 다른 정보기술에 비해서 그 필요성이나 인력이 많이 줄어들고 있는 시점임. 한국어 정보처리 환경을 개방형으로 전환하고 연구기관의 지속적인 개방형 연구 환경을 통하여 한국어 정보처리 관련 전문 기술 인력을 확보할 수 있음.
- **한국어 자원 활용 정보처리 기술 선도:** 한국어 정보처리 모듈에 대한 광범위한 지식 라이브러리를 바탕으로 다양한 테스트 업무를 수행하고, 이를 통하여 가장 높은 정확도와 효율성을 자랑하는 한국어 정보처리 기술을 개발함으로써 관련 기술에 대한 효율성을 극대화할 수 있음.
- **즉각적인 언어 현상의 반영:** 현재 한국어에서 발생되고 있는 다양한 언어 현상, 혹은 번역 메모리상으로 기술 가능한 구어체 번역 정보, 뉴스 정보에서 제공되는 실시간 동영상 전사 말뭉치 지식 등을 공유를 통하여 나타내고, 실시간 언어 정보를 나타냄으로써 현재 실제 사용되고 있는 언어 현상을 반영하고, 언어/용어의 생명 주기를 관리할 수 있는 프레임워크를 마련할 수 있음.

○ **학교/기업의 효율성 향상**

- **한국어 정보 관련 중복투자 방지:** 이미 구축된 말뭉치나 사전 정보를 각 정부기관/연구기관/기업별로 만들어서 한국어 정보에 관련된 중복투자가 일어날 가능성이 줄어듦. 아울러 비슷한 기능의 컴포넌트를 하나의 통합된 환경에서 제공함으로써 한국어 관련 기술 개발에 선택과 집중 및 중복투자를 미연에 방지할 수 있음.
- **각 학교/기업별 기술 전문화: 기업별로** 관리 기술, 분석 기술, 평가 기술 등을 특화하여 개발하고, 상업적으로 활용할 용례를 기업들이 활발하게 개발함으로써 한국어 전체의 정보 효율성을 극대화할 수 있음.
- **학교/기업의 국어 자원 활용 서비스/솔루션 개발 편의성 증대:** 대학에서는 한국어 관련 핵심 기술을 가지고 있지 못함으로써 해결할 수 없던 여러 가지 문제를 공유 자원을 통해서 쉽게 해결할 수 있게 됨. 이는 기업에서도 언어 처리 및 기반 기술에 소요되는 비용을 상당부분 절감할 수 있음.

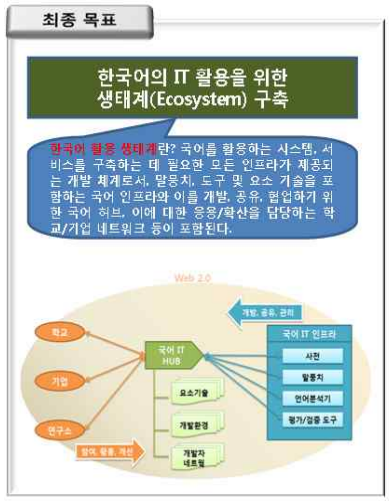
○ **사회 문화적 이익**

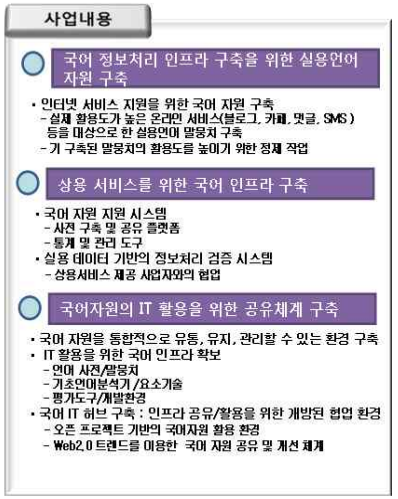
- **한국어 자원에 대한 관심 유발 :** 지금까지 사회적으로 소홀이 되어 온 한국어 자원에 대한 일반 대중의 관심을 유발할 수 있음.
- **한국어의 국제화 이바지:** 다국어 정보들을 통합적으로 관리할 수 있는 오픈 프레임워크를 제공함으로써, 한국어가 외국어와 함께 통합적인 환경에서 연동될 수 있는 기반을 마련하게 됨.

1.2. 사업 내용

○ 한국어 활용생태계(ecosystem)란 국어를 활용하는 시스템 및 서비스를 구축하는 데 필요한 모든 인프라가 제공되는 개발 체계로서, 말뭉치, 도구 및 요소 기술을 포함하는 국어 인프라와 이를 개발, 공유, 협업하기 위한 국어 정보 허브(hub), 이에 대한 응용/확산을 담당하는 학교/기업 네트워크 등을 포함하고 있음.

○ 본 사업에서는 이러한 한국어의 IT 활용을 위한 생태계를 구축하는 것을 목표로 함.





1.2.1. 한국어 정보처리 인프라 구축을 위한 실용언어 자원의 구축

1.2.1.1. 사업 목표

○ 한국어 정보처리를 위한 실용언어 인프라를 구축함.

○ 아울러 기구축 말뭉치에 대한 정보를 추출하여 재사용성 있게 가공하며, 이를 운영할 수 있는 언어처리 인프라를 구축함.

(1) 실용언어 자원의 구축

○ 인터넷 관련 IT 기업에서 시의적절한 실용언어 처리 문제를 해결하는 데 필요한 기초 자료를 제공하기에 적절한 규모의 실용언어 자원을 구축함.

○ '블로그', '카페', '댓글', '상품평', 'SMS' 등 실용언어 말뭉치를 300만 어절 규모로 구축함.

○ 초기 구축은 인터넷 커뮤니티에 등록된 최신 3년 이내 자료들을 대상으로 하며, 이후 각 개발년도에 새로 생성된 데이터가 반영하여 최신 데이터가 유지될 수 있는 프로세스를 마련함.

○ '댓글', 'SMS' 자료는 최소 3어절 이상인 문서들을 수집함.

○ 철자 오류, 띄어쓰기 오류는 수정하지 않고, 오류 태그를 유형별로 분류하여 부여함. 단, 띄어쓰기 오류가 3어절 이상인 것은 띄어쓰기 오류를 수정함.

○ 말뭉치를 직접 활용할 인터넷 IT 기업들의 실제 수요를 최대한 반영하여 실용적 가치를 고려하여 구축함.

(2) 기구축 말뭉치의 정보 추출 및 가공

○ 형태 분석 말뭉치에서 추출될 수 있는 각종 정보 자료 유형들을 조사하고 정보 추출 도구를 개발함.

○ 기구축된 말뭉치에 대해 '토큰 단위'의 품사 태깅 말뭉치로 변환 및 가공하여 정보검색 분야에서 각종 응용 소프트웨어를 개발하는데 실질적으로 필요한 'bag of words' 형태의 말뭉치를 구축함.

○ 다양한 형태의 정보유형들에 대해 각 자료에 대한 통계자료 생성 도구를 개발하여 말뭉치 사용자들에게 배포 가능하도록 함.

○ 배포된 정보 추출 및 가공 도구를 말뭉치 사용자들이 편리하게 활용할 수 있도록 함.

○ 복합명사의 분해-결합, 접두/접미사의 분해-결합 등 일관성 혹은 일정 기준을 유지해야 하는 정보자료의 유형을 인지하고 분해 기준의 타당성에 따라 오류를 수정함.

○ 말뭉치에 나타나는 다양한 언어 현상에 대한 각종 분석 도구를 개발하여 관련 분야에 보급함.

(3) 국어 기초자원의 활용 및 언어처리 인프라 구축

○ 한국어 정보처리 관련 연구-개발자 및 유관 기관에서 공통적으로 필요한 소프트웨어의 신뢰도 및 성능을 제고함.

○ 한국어 정보처리, 정보검색 분야의 신진 연구자들이 이 분야의 지식을 습득하고 학습하는 과정에서 필요한 기본적인 소프트웨어들을 제공하여 이 사업의 홍보 효과를 최대화하도록 함.

○ 관련 분야 연구-개발자 및 정보검색 소프트웨어 사용자들이 필요한 국어 기초자원을 사용자 관점에서 재구축, 가공하여 배포하여 이 사업의 결과물이 최대한 많은 사용자들에게 도움이 될 수 있도록 함.

○ 유니코드 관련 문제, 대용량 말뭉치의 빈도 계산 속도 문제 등 다수의 말뭉치 사용자들에게 공통적으로 발생하는 문제를 해결해 주는 소프트웨어를 개발하여 open source 형태로 제공함.

○ 기구축 및 구축 예정 말뭉치의 활용가치를 극대화하고, 국어정보화 기술의 우위를 확보함.

1.2.1.2. 세부 사업 내용

(1) 실용언어 자원의 구축

기구축된 말뭉치와 별개로 실용적인 언어생활과 관련된 아래 분야들에 대해 '토큰 단위'의 품사 태깅 말뭉치를 구축하며 최신 데이터가 유지될 수 있

는 프로세스를 마련함.

○ 블로그 말뭉치: 100만 어절

○ 카페 말뭉치: 100만 어절

○ 댓글 말뭉치: 50만 어절

○ 상품평 말뭉치: 30만 어절

○ SMS 말뭉치: 20만 어절

○ 실용 말뭉치에 적합한 품사 표지(part-of-speech tag)를 정의함.

○ 실용 말뭉치 수집 및 정제, 가공하는 과정에서 기구축된 기초 말뭉치와 달리 댓글, 상품평, SMS 말뭉치는 수집하는데 애로상황이 발생할 수 있음.

○ 각 개발년도마다 최근의 데이터가 말뭉치에 반영될 수 있도록, 업계와의 유기적인 협조를 포함한 최신 데이터 반영 프로세스를 구축함.

○ 말뭉치 구축 및 관리 도구 개발: 구축 및 활용의 편의를 위한 관리도구

○ 말뭉치 구축과정에서 발생할 수 있는 오류들을 탐지하는 검증 도구를 개발하여 오류 발생을 최소화하도록 함.

(2) 기구축 말뭉치의 정보 추출 및 가공

○ 기구축 말뭉치에 수록된 다양한 정보를 다수 사용자가 쉽게 활용할 수 있도록 공통적으로 활용되는 정보들을 추출하여 연구-개발자들이 실용적으로 필요한 정보만을 쉽고 편리하게 사용할 수 있도록 함.

○ 검색엔진에서 색인의 단위, 기타 응용 분야에서 정보전달의 최소 단위가 되는 '토큰 단위' 정보 등 추가로 가공이 필요한 것은 기구축 말뭉치로부터 '토큰 단위' 말뭉치로 변환함.

○ 형태소 단위 품사 표지를 참조하고 정보 전달의 실질적, 효율적 단위를 충분히 고려하여 정보검색 응용 소프트웨어의 성능을 최대화할 수 있도록 '토큰 단위'의 품사 표지를 다시 정의함.

○ 형태론적 변형이 유지된 토큰의 인식 단위를 정의해야 하며, 이 토큰은 정보처리 분야에서 의미가 있는 토큰이어야 함.

○ 형태분석 말뭉치 6천만 어절에 대한 '토큰 단위' 품사 태깅 말뭉치 구축

○ 형태의미 분석 말뭉치 1천만 어절에 대한 '토큰 단위' 품사 태깅 말뭉치 구축

○ 기구축 말뭉치에서 품사 유형별 정보, 품사별 형태소 정보 등 다양한 유형들에 대한 언어정보를 추출하여 제공함. 말뭉치에서 많은 사용자들이 공통으로 필요로 하게 되는 정보들이 있으며 이를 미리 잘 정리하여 웹사이트 등에 제공함.

(3) 국어 기초자원의 활용 및 언어처리 인프라 구축

○ 한글 처리를 위해 개인 연구-개발자 및 관련 기관에서 공통적으로 필요한 도구 및 기초 자원 활용 도구를 개발하여 보급함. 신뢰도가 높고, 성능이 우수한, 완성도가 높은 유용한 도구를 개발함.

○ 언어자원 활용에 필요한 언어처리 기초 도구: 미등록어 수집 도구, 언어자원 통계처리 도구, 유니코드 변환 등

○ 말뭉치 정제 도구: 띄어쓰기 오류 교정, 철자오류어 수정 등

○ 말뭉치에서 추출한 기초 언어처리용 전자사전

○ 한글 코드 관련 자료 및 도구, 코드변환표 등

○ 다수의 국어 사용자들이 혼동하는 띄어쓰기 오류 관련 자원 및 수정 도구

○ 다수의 국어 사용자들이 혼동하는 철자 오류 관련 자원 및 교정 도구

○ 유니코드 문서 처리 도구

- 유니코드-KS완성형 변환, 한자-한글 변환 등

- 문자 유형별 태깅 도구: 한글/한자/일본어/특수문자 등

○ 외국인을 위한 한글학습 보조 시스템: 형태소 분석 및 사전 탐색

○ 다수 사용자의 한글 사용 편의성 높이기 위하여 기본적인 한글 정보처리 도구와 관련하여 발생하는 분쟁 문제를 해결하는 방안을 모색함. 그 예로, 한-영 자동전환, 웹 브라우저의 'WWW/ㅈㅈㅈ' 입력모드 자동 전환, 질의어 자동완성 기능, 핸드폰 문자입력 방식 등 한글 관련 특허 문제에 대한 해결 방안 혹은 새로운 접근 방안을 모색함.

1.2.2. 언어 산업과 연계한 언어 인프라 구축

1.2.2.1. 사업의 목표

한국어의 상용 서비스를 위한 기반 언어 인프라를 구축하고, 이를 위해 용어 사전 구축 및 실용 데이터 기반 색인 및 검색 평가 세트를 구축함.

1.2.2.2. 세부 사업 내용

○ 포털 사업에 있어서 정보검색 및 기계번역 등의 중요성이 높아지고, UCC의 활성화로 이를 가공하기 위한 기술이 요구됨에 따라서 언어자원의 활용 분야도 점차 다양해지고 있음.

○ 인터넷 고객들이 가장 많이 사용하는 언어처리 서비스가 정보검색이므로 정보검색에서 요구되는 정보검색 인프라와 정보검색 서비스를 고려해야 할 것임.

○ 정보검색 인프라의 경우는 업체별로 산재하여 구축하고 있는 색인기술과 전자사전을 공유함으로써 중복투자를 최소화하는 효과를 기대할 수 있을 것임.

○ 공신력 있는 테스트 Set의 구축/공유를 통해 국내 정보검색 서비스의 전체적인 품질 수준을 진단하고, 또 품질의 향상을 촉발하는 효과를 기대할 수 있음.

○ 앞으로 새로운 검색 서비스의 출현을 지원하기 위해 주제별로 분류된 말뭉치를 구축하여 공유함으로써 검색 산업의 발전에 기여할 수 있을 것임.

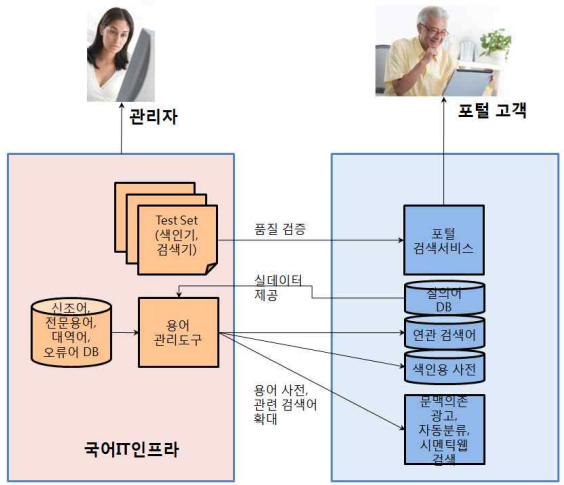
○ 이러한 인프라 외에 정보검색 서비스 및 다양한 응용 서비스를 구축함에 있어서 필요로 하는 언어자원과 구축 도구에 대한 필요성도 증가하고 있음.

○ 비정형 문서에 대한 가공 기술의 중요성이 증대되고 있어 이에 대한 분석/필터링 기술의 개발이 시급히 요구되고 있음.

○ 이러한 검색 및 응용 서비스의 운영에 있어서 지속적으로 신조어, 전문용어, 대역어, 심지어 오류어 패턴의 구축이 필요함.

○ 이렇게 확보된 언어 자원을 산업체에서 활용할 수 있도록 함으로써 중복 투자로 인한 낭비를 줄일 수 있을 것임.

○ 또한, 이러한 언어자원 및 기초모듈, 구축도구 등은 집단지성을 활용할 수 있도록 개방되고 공유되며, 개선될 수 있는 체제로 구성되어 다양한 업계의 전문가들을 통해 지속적으로 업그레이드되는 방향으로 진행되어야 할 것임.



<그림 9> 국어 IT 인프라의 포털 서비스 활용

○ 국어 IT 인프라의 실효성을 위해서는 포털 사이트로부터 축적된 실데이터를 기초 자료로 활용할 필요가 있음.

○ 포털 사이트의 질의어 DB, 웹크롤러를 통해 수집된 블로그 포스트, 각종 커뮤니티 서비스에 게시되는 게시물 등이 국어 IT 인프라의 기초 자료가 될 것임.

○ 이를 정기적으로 제공받아 국어 IT 인프라에서는 각종 용어 데이터 및 용어 관리 도구를 제공하고, 또 기초 자료를 정제하여 색인기와 검색기의 품질 테스트를 위한 정제된 평가 세트를 제공함.

○ 포털 사이트 사업자는 이러한 정보를 Open API 등의 방법으로 활용하여 서비스를 개선하고 그 결과를 다시 국어 IT 인프라로 제공하는 순환 싸이클을 구성함.

○ 제공된 언어 자원은 포털 사업자들의 차세대 서비스인 "문맥의존형 광고", "자동분류 및 콘텐츠 필터링", "시맨틱웹검색" 등의 기초적인 언어 자원으로도 활용 가능함.

(1) 용어 사전 구축 및 공유 플랫폼 구축

○ 신조어, 전문용어, 대역어, 오류어 등 실용 분야에 대한 용어 사전 구축

○ 상용 서비스에 활용 가능한 수준의 용어 규모

○ 웹기반의 용어 사전 관리/공유 환경 및 API 제공

(2) 실용 데이터 기반의 색인 및 검색 Test Set 구축

○ 포털 사업자로부터 실용 데이터를 지속 공급받는 체계 구축

○ 색인기 품질 테스트용 Test Set 구축 및 업그레이드 체계 마련

○ 검색기 품질 테스트용 Test Set 구축 및 업그레이드 체계 마련

1.2.3. 국어자원의 IT 활용을 위한 공유 체계 구축

1.2.3.1. 사업 목표

한국어 정보를 통합적으로 유통하고 이를 중심에서 유지, 관리할 수 있는 통합화 된 환경을 구축함.

1.2.3.2. 세부 사업

(1) 국어 IT 인프라 설계

○ 국어 IT 정보는 단순히 하나의 허브를 구축하는 것만으로는 구축될 수가 없음.

○ 집합적인 지식을 모으는 방식도 설계할 수 있겠지만, 국가 연구기관 사이의 협약을 통하여 국가기관 내 전산센터와 협력을 통하여 국어 정보를 실질적으로 수집하고, 이를 라이브러리로 모을 수 있는 기반 구조를 만들 수 있어야 할 것임.

○ 이는 인터넷 포털이나 기업, 국책 연구기관도 비슷한 방식으로 하여 국가의 언어 자원을 통합적으로 수집할 수 있음.

○ 수집되고 공유된 정보는 국어 IT 인프라 환경을 구성하는 데 큰 역할을 담당함.

○ 본 구조는 각 연구기관 및 정부기관과 유기적인 협의 및 협조를 통하여 구축되어야 할 것임.

(2) 국어정보화 IT 허브 구축

○ **소스 인프라 구축**

- 일반 사용자들이 구축한 소스를 업로드 하거나 다운로드 하고, 이에 대한 사용자들의 의견을 물어보거나 디버깅을 하여 통합적으로 버전을 관리할 수 있는 SourceForge 개념의 소스 관리 인프라를 구축.

○ **언어자원 관리 인프라 구축**

- 물리적 데이터베이스나 파일 시스템에서 언어자원 인프라와 함께 통신할 수 있는 시스템 기반을 구축.

○ **라이브러리/사전 인프라 구축**

- 구조화된 라이브러리, 표준화된 언어 분석 라이브러리를 기반 인프라를 구축한다. 이를 통하여 기본적인 언어 분석 및 사전 처리 작업이 구성

○ **테스트 인프라 구축**

- 테스트 정보를 관리하고 검색하며, 새롭게 구축한 루틴을 서버에 등록하여 실제 언어 자원에 적용해 보고, 이의 결과를 확인할 수 있는 엔진, 이를 관리하는 서버 세트를 구성.

○ **공유 인프라 구축**

- Open API 혹은 컴포넌트 형태로 배포 가능한 서비스, 시스템을 관리하는 인프라를 구축함.

○ **의사결정 인프라 구축**

- 사용자들의 의견을 통합하여 새로운 컴포넌트를 요소 기술로 적용할 것인지, 언어자원에서 각 규칙의 타당성을 판단하거나, 프레임워크의 설계 결과를 적용할 것인지 등 핵심적인 의견을 교환하고 토론하는 의사결정 인프라를 구축.

○ **프레임워크 설계 인프라 구축**

- 한국어 정보 처리도 형태소 분석이나 문형 분석/문형 추출/태깅과 같은 기반 컴포넌트와 응용 컴포넌트로 구분될 수 있음.
- 개방형 환경이라 하더라도 가장 기반이 되는 컴포넌트 설계 환경은 기반으로 제공되어야 함을 의미.
- 프레임워크 설계 인프라는 이러한 프로그래밍 인터페이스, 컴포넌트 구조를 설계할 수 있는 환경을 제공.

○ **버전 관리 인프라 구축**

- 통합적으로 각 정보들의 버전을 통합적으로 관리할 수 있는 CVS 형태의 환경을 제공.
- 보다 광의의 개념으로 적용하여서 각 수정된 내용에 대한 수정, 복원 방법을 효율적으로 제공하여서 버전 관리가 유기적으로 관리되고, 실제 엔드 사용자에게 제공되는 배포 버전을 관리할 수 있는 기반을 마련함.

(3) 실용언어 자원 인프라 마련

○ 지금까지 세종계획에서 구축되었던 말뭉치는 현재 온라인상에서 이루어지고 있는 다양한 언어생활을 반영하고 있지 못함.

○ 실용적인 언어자원 인프라를 관리하고 운영하기 위한 인프라를 마련해야 함.

○ **버전 관리 시스템**

- 각 등록된 말뭉치의 중복 여부, 갱신 여부 등을 실시간으로 관리하고, 이를 사용자들의 참여를 통해서 운영할 수 있도록 해야 할 것임.

○ **사용자 등록 말뭉치 환경**

- 19세기 한국어, 전문 분야의 학술 문서 등의 정보는 온라인 상으로 수집되거나 폼 형식으로 모여질 수 없음.
- 특수 정보들은 과거에도 다양한 형태로 큰 비용을 들여 구축될 수 있었지만, 연구 목적의 성격이 강한 말뭉치 정보들은 실질적으로 사용자들이 이를 업로드 형식으로 파일을 구축하고 등록할 수 있는 온라인 기반 환경을 제공해 주어야 할 것임.

○ **온라인 기반 말뭉치(블로그 말뭉치, 카페 말뭉치, 댓글 말뭉치)**

- 온라인 기반 말뭉치는 최근 온라인에서 많은 누리꾼들에 의해 작성되고 있는 블로그, 그리고 동호회나 카페에서 작성되고 있는 수많은 페이지의 정보를 포함함.
- 이와 함께 실시간으로 계속 증가하는 댓글의 경우 그 시사하는 바가 매우 큼.
- 사회적인 이슈가 되고 있는 악플이나 그때그때 만들어지는 신조어들에 대한 정보를 포함하고 있기 때문에 현재의 언어현상에 대하여 가장 정확한 정보를 제공할 수 있음.
- 각 정보를 실시간으로 수집하거나, 라이브러리 형태로 구축할 수 있도록 기반 환경을 마련함.

○ **대화체 말뭉치(SMS 말뭉치, 대화 말뭉치, 음성 연동 말뭉치)**

- 대화체 말뭉치는 수집되기 어려운 부분은 있으나 SMS로 주고 받는 말뭉치, 그리고 일반인들이 이야기 하는 대화 말뭉치로 나눌 수 있음.
- 음성 인식 및 합성 시스템을 위한 음성 연동 말뭉치의 경우, 최근 방송사에서 제공되는 모든 동영상에 대해서는 문자 전사 작업이 함께 이루어지고 있는데, 이를 기술적으로 연동하여 말뭉치로 구축할 경우 향후 언어 정보 처리에 큰 도움을 줄 수 있음.

(4) 언어도구 및 요소기술 평가 인프라 마련

○ 언어 도구 및 요소기술이 평가되기 위해서는 하나의 표준화된 평가 방법이 구조화된 문서로 지원될 수 있어야 함. 아울러 언어 도구에 의해서 정확한 절차에 의해 분석이 이루어질 수 있어야 할 것이며 그 결과를 쉽게 확인할 수 있는 시각화 환경이 함께 마련되어야 할 것임.

○ **형태소/구문/의미/화행/개체명 분석 평가데이터 및 평가 지원 도구**

- 언어 도구 및 요소기술을 평가하기 위해서는 하나의 표준화된 시스템에 이를 입력하고 분석할 수 있어야 함.
- 형태소와 구문/의미/화행/개체명 분석 정보는 기존 라이브러리 정보를 하나의 세트로 입력을 하고, 새롭게 테스트 하는 정보를 입력하여 실제 규칙 정보에 맞는 정보, 중의성이 있는 정보, 분석이 되지 않는 정보로 각각 돌려주며, 특히 분석되지 않는 정보에 대해서는 미등록 개체명이나 새로운 패턴으로 분석될 수 있는지 지원이 가능토록 시스템을 구성해야 함.

○ **정보검색/추출/분류 평가데이터 및 평가 지원 도구**

- 정보 검색 결과에 따른 사용자의 만족도, 추출이 정확하게 되었는지, 코드상으로 잘 못된 문자들이 있는지 다양한 형태의 정보들을 관리하고 이를 평가하는 도구들이 포함될 것임.
- 자동 분류 환경에 적합한 신문, 블로그, 전문 정보 등을 포함하며, 타 연구과제에서 연구되는 평가 기준을 준용하여 평가 데이터를 통합 관리하고 분석하는 환경을 구성해야 함.

○ **통합 언어정보 관리도구**

- 언어 자원은 통합적으로 관리되어야 함.
- 지금까지 세종계획의 정보는 각 요소들이 어휘/형태소와 의미번호 형식으로 연결되어 있었으나, 통합적으로 관리되지 않았기 때문에 이 사이의 연관성이 정합성 없이 관리되어 왔음.
- 실질적으로 대규모 언어 자원이 제대로 구축되고 운영되기 위해서는 먼저 통합적인 언어정보 관리 도구가 구축될 수 있어야 하며, 이를 통해 실질적으로 사전의 어휘와 말뭉치, 온톨로지나 어휘의미망 지식들이 하나의 유기적인 체제 아래에서 관리될 수 있음.

(5) 오픈 프로젝트 기반 국어자원 활용 환경

○ 기초 언어 분석기는 모든 정보를 통합하여 오픈된 환경에서 유기적으로 운영될 수 있는 기본 플랫폼을 제공하는 것을 목표로 함.

○ 국어 자원을 통합적으로 활용하고 응용하기 위한 기반 데이터를 마련하고 그 환경을 구축하는 데 있음.

○ **오픈 프로젝트를 위한 기초 언어 분석기 개발 또는 확보**

- 기초 국어자원 활용을 위한 응용 컴포넌트를 개발하여 이를 외부에 공개하고, 각종 모듈을 통하여 사용자들이 분석한 결과가 다시 개발 과정에 반영되는 형식으로 새로운 개념의 응용 모듈을 만들어야 함.
- 핵심 엔진 개발과 분석용 라이브러리 구조를 함께 공개하고, 이를 통하여 분석 및 활용 기술이 진일보할 수 있는 기반을 마련해야 함.

○ **오픈 프로젝트를 위한 공유사이트 개설 또는 확보 및 관리, 라이선스 정책 수립**

- 오픈 프로젝트는 하나의 정책이자 인간의 집합적인 지식에 의하여 그

중의성이나 분석 결과가 결정되도록 구성되어야 하므로, 하나의 공유된 사이트를 개설하여 사람들의 집합지식이 반영될 수 있도록 구성되어야 함.

- 구글의 Knol처럼 특정 글을 올리는 사용자에 대해서 어떤 강한 Authority를 주며, 언어 분석 결과에 대해서 판단하고 분석 방향을 결정할 수 있는 관리 방법론 및 환경을 마련함.
- 공유 정보에 대한 라이선스 정책을 수립함. 통상 라이선스 정책은 GPL(GNU General Public License), LGPL(GNU Lesser General Public License), BSD, 아파치 라이선스 정책이 있는데, 기업에서의 활용을 위해 LGPL 라이선스 방식을 기준으로 하여 수립해야 함.

○ **기술 표준화**

- 기술 표준화는 이미 ISO/TC37 위원회에 참석하는 것 이외에도 앞서 언급한 LISA 및 OASIS와 같은 국외 말뭉치 관련 학회 등 ISO나 W3에 진입하고자 하는 많은 기업 기반 연구회들에도 적극 참여하여 세종계획에서 구축된 수많은 성과물이 실질적으로 응용될 수 있도록 함.

(6) 국가 기반 언어정보 서비스 허브 구축

○ 클라우드 컴퓨팅 개념을 도입하여 인터넷 상으로 모든 언어정보를 공유와 협업에 의해 구축한다고 하더라도 궁극적으로는 하나의 통제되는 조절점이 있어야 함.

○ 국가의 언어 정보 역시 한민족의 문화 정체성과 직결되어 있으며, 이는 오랜 역사적 발전 과정을 보아서도 증명되고 있음.

○ 만약 클라우드 컴퓨팅, web 2.0 개념에 의한 공유와 협업에 의해 가중치가 결정되는 시스템이 구축되어도 언어 정책에 있어서 흔들리지 않고 바뀌지 않아야 하는 하나의 뿌리 같은 환경은 필수적으로 운영되어야 함.

○ 국가 기반 언어정보 서비스 허브는 이러한 최종 한국어 정보를 통합 저장하고, 중의성이 있는 언어지식에 대해서 최종적인 판단을 내리는 핵심 역할을 담당할 것임.

○ 온톨로지 정보, 사전과 전문 정보에 대한 라이브러리 정보, 구축된 말뭉치를 제공하는 서비스를 포함함.

○ "기초 언어 분석기 오픈 프로젝트"와 함께 연동하여 점차 그 라이브러리

가 확보되고 통합되며, 나아가 이 서비스 역시 오픈 프로젝트화 될 수 있음.

○ **국가 기반 온톨로지 정보의 표준화 및 서비스 인터페이스 구축**

- 국가 기반 온톨로지 정보는 일반론적인 의미의 최상위 의미를 가지는 온톨로지 지식(5000개~1만개) 수준의 정보를 핵심적으로 정부에서 관리하는 것임.
- 이 최상위 개념체를 바탕으로 관련 도메인이나 해당 관련 지식들의 정보를 풍부히 만들 수 있으며, 모든 관련된 온톨로지 지식은 이 정보에 대한 하나 이상의 relation 정보를 가질 수 있도록 함.
- 아울러 정부의 관련 유관 기관의 온톨로지 정보와도 연계를 맺을 수 있도록 구성함.

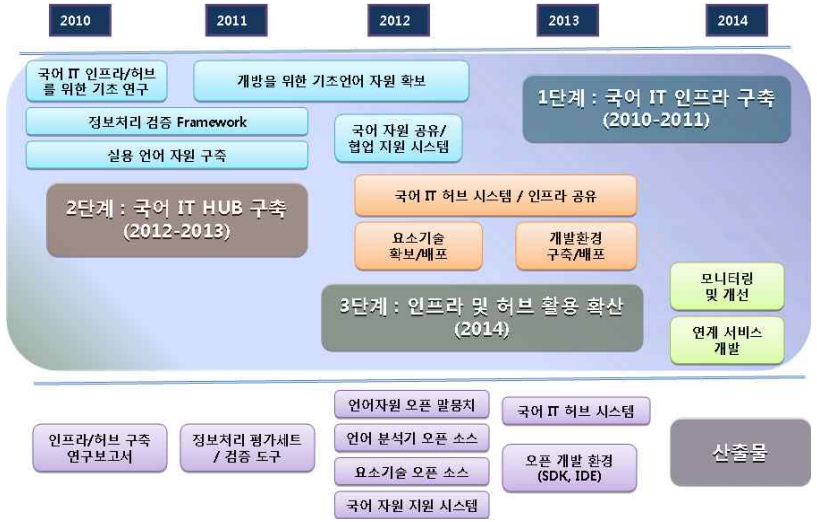
○ **국가 기반 전문 정보 및 언어 라이브러리 서비스 IT 허브 구축**

- 표준국어대사전 및 세종전자사전, 그리고 이와 연관된 각종 분석 언어 정보를 라이브러리 형태로 구축하여 사용자에게 서비스 함.
- 이를 가능케 하는 것은 기존에 구축된 방대한 분량의 세종계획 라이브러리뿐만 아니라, 지속적으로 구축된 전문용어 사전, 아울러 이에 대한 의미 태깅, 형태소 분석 태깅 정보가 존재하고 있기 때문임.
- 확장 및 관리를 위한 인터페이스를 통하여 이러한 정보는 효율적으로 관리되고 확장됨.

○ **국가 기반 언어정보 서비스 허브 구축**

- 국가에서 수집된 다양한 관보, 문서 정보들을 언어정보 처리가 가능하도록 기초 언어 분석기 오픈 프로젝트 서비스와 연동될 수 있도록 구성함.
- 기초 언어 분석기 서비스는 각 분석 요소에 대한 정보를 제공하며, 본 서비스에서는 국가의 기반 언어정보에 대한 URI와 실제 원문을 가지고 있도록 함.
- 언어 정보에 대한 통합적이고도 직관적인 정보 관리가 가능하게 되며, 내부적으로 버전이 변경되는 경우에도 일관되게 정보 체계를 관리할 수 있음.

1.3. 단계별 로드맵



2. 미래를 준비하는 한국어

2.1 사업의 필요성

○ 세계인의 삶의 환경과 방식을 급격히 변화시키고 있는 인터넷은 앞으로도 끊임없이 발전하고 진화해 나갈 것으로 기대됨.

○ 인터넷이 멀티미디어 방식으로 다변화되고 있음에도 현재로는 언어 텍스트 정보가 대부분을 차지하고 있고 앞으로도 문자정보의 중요성은 결코 줄어들지 않을 것으로 보임.

○ 인터넷의 진화의 핵심 방향은 "의미망(Semantic Web)"이라고 이미 인터넷 창시자인 Tim Berners-Lee에 의해서도 제시된 바 있음 (Berners-Lee et al. 2002). 여기에서 '의미'는 언어의 의미와 관련된 현상을 포괄적으로 지칭하는 것으로 이해할 수 있음.

○ 언어는 본질적으로 형태(문자나 소리)와 의미의 결합체라는 점에서 볼 때 초기 및 현재의 인터넷 검색은 기술의 한계로 형태에 치중할 수밖에 없었고, 그 결과 의미적으로 연관성이 없는 수많은 '쓰레기' 정보까지 찾아 주거나, 아니면 의미적으로는 직접 연관성이 있는데도 형태적으로 차이가 나는 표현은 거의 찾기가 어려웠음.

○ 궁극적으로는 사람처럼 말하고 알아들을 수 있는 인공지능이 개발되어 인터넷 검색을 할 수 있을지는 몰라도, 실제로 현실적이고 단계적인 방안으로 두 가지 방식으로 나누어 볼 수 있음. 첫 번째는 언어의 의미적 특성을 체계적으로 파악하여 그러한 정보를 웹자료에 직접 반영하거나 연계함으로써 의미검색을 가능하게 하는 방안이고, 두 번째는 의미적인 웹자료의 생성에 집단지성형태로 모두 함께 참여하도록 유도하는 방안으로 "Web 2.0"이라는 참여형 웹의 방향을 예로 들 수 있음. 이중에서 후자는 주로 인터넷 기술의 문제로 부각되고 있고, 전자는 언어에 대한 보다 근본적인 연구 및 정리가 선행되어야 한다는 점에 공감하고 있음.

○ 영어나 기타 유럽어, 일본어, 중국어 등 언어별로 의미적 관점에서 그 언어의 특성을 파악하려고 적극적으로 노력하고 있고, 한국어에 대한 같은 차원의 연구도 현재 이루어지고 있음.

○ 그런데, 그러한 노력의 바탕이 되는 대규모 자료, 즉 말뭉치는 개인적인 차원에서는 구축이 거의 불가능함. 인력(고급지식을 가진 전문가), 비용, 시간이 매우 많이 드는 작업으로 중소규모의 언어처리 산업계나 대학별 연구실에서 감당하기 어려움.

○ 또한 이미 각 연구실이나 업체에서 개별적으로 분산되어 개발 중인 자원들을 통합하고, 또 더 많은 연구자나 업체가 활용할 수 있는 형태로 만들어줄 필요성이 크게 대두되고 있음.

○ 따라서 이처럼 의미적 특성 파악에 핵심이 되는 언어자원을 통합 구축하고 그러한 자원을 활용할 수 있도록 적극 유도하는 것은 한국 및 한국어의 미래를 위한 중요한 국가적 사업임.

○ 가까운 미래를 위해 현재 자원 구축 및 적극적인 연구가 필요한 의미관련 주요 분야로, 현재 전 세계적으로 많이 연구되고 있고 또한 국제표준기구(ISO)에서도 자원구축 방식에 대한 표준화 작업이 진행 중인 분야임.

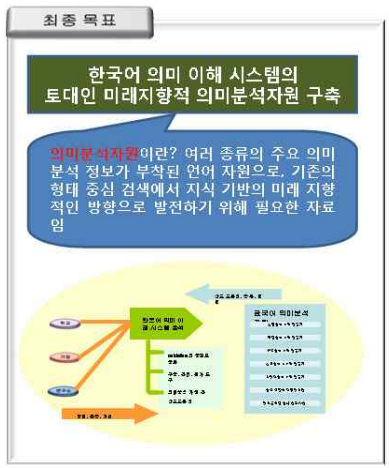
○ 국어정보화 1단계 사업이라고 할 수 있는 '21세기 세종계획'의 결과물을 2단계 사업에서 효율적으로 활용·응용하게 함으로써 국가지원 사업의 연속성 확보

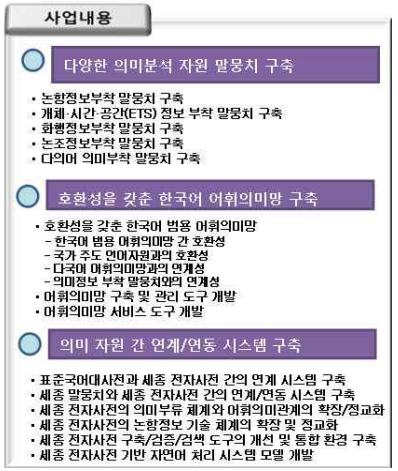
○ 규모와 내용면에서 세계적 수준의 본격 전자사전인 세종전자사전의 각종 정보를 특수 응용 목적의 언어자원 구축에 적극 활용하게 함으로써 관련 기술개발의 시간 단축 및 투자비용 절감 효과 유발

2.2. 사업 내용

○ 의미분석 자원이란 여러 종류의 주요 의미분석 정보가 부착된 언어 자원으로, 기존의 형태 중심 검색에서 지식 기반의 미래 지향적인 방향으로 발전하기 위해 필요한 자료임.

○ 미래를 준비하는 한국어 부문 사업은 이러한 의미분석 자원을 구축함으로써 한국어 의미 이해 시스템의 토대를 마련하는 것을 목표로 한다.





2.2.1. 다양한 의미분석 자원 구축

2.2.1.1. 논항정보부착 말뭉치 구축

(1) 사업 내용 정의

○ 본 사업의 주요 내용은 문장의 의미적 특성을 파악하는 데 필수적인 논항구조를 자동으로 파악하는 데 사용하기 위해 한국어 논항구조 말뭉치를 구축하는 것임.

○ 구문분석과의 연계성을 위해 세종계획에서 구축한 구문분석말뭉치 (80만 어절 규모)를 바탕으로 하여 필요한 정보를 추가하는 방식으로 함.

- 추가될 필요 정보로는 생략논항, 논항 구조정보 등이 있음. (Penn Korean Treebank 방식 참조).
- 구어의 논항구조정보를 위해 세종 구어 형태소 분석 말뭉치에서 20만 어절 규모의 자료를 선택하여 구문분석을 하고 논항정보를 부착.
- 태그정보 부착을 위한 말뭉치 구축, 검증, 활용 도구 개발

(2) 사업 목표

○ 향후 5년간 논항정보가 부착된 총 1백만 어절 규모의 현대 한국어 말뭉치 구축

- 그 중 80만 어절은 세종 구문분석 말뭉치를 대상으로 그 말뭉치를 정제 및 보완 후에 논항정보 부착.
- 나머지 20만 어절은 세종 구어 말뭉치에서 선별하여 선택한 후 필요한 구문 분석과 아울러 논항정보 태그 부착.

○ 범용성 높은 표준적 태그 설정 (세종말뭉치와의 호환성 고려)

○ 태그정보 부착을 위한 말뭉치 구축, 검증도구 개발

- 구문분석 말뭉치 구축 도구 (구어말뭉치)
- 논항구조 말뭉치 구축도구
- 논항구조 말뭉치 검증도구

○ 논항정보 부착 말뭉치 검색, 활용도구 개발

- 논항구조 말뭉치로부터 형태소, 구문, 논항 관련 정보 추출 도구

(3) 세부 사업

○ 논항정보 부착 말뭉치 주석 방식에 대한 국내외적 기준 조사 및 기준 확립

- 국제적 기준에 대한 포괄적 조사
- 국내 관련 학계의 용어 및 분류 기준 등 조사

○ 기존 한국어 자원과의 연계 방안 연구

- 세종전자사전/표준국어사전에서의 논항 정보 검토 및 연계 방안 연구
- 어휘망작업과의 연계 방안 연구

○ 시험 말뭉치 구축: 5만 어절 규모

- 태깅 도구 개발 및 운용 시험
- 태깅 검증 도구 개발 및 시험 구축된 말뭉치 검증

○ 말뭉치 목록 선정

- 세종구문분석말뭉치 80만 어절 포함
- 세종구어말뭉치에서 20만 어절 추가: 균형성 및 다양성 문제 검토

○ 말뭉치 본격 구축

○ 말뭉치 활용 방안 수립

- 활용 방향 정리 및 기초 활용 도구 개발
- 말뭉치 활용방법 홍보
- 사용자 의견 수렴 및 반영

2.2.1.2. 개체·시간·공간(ETS) 정보 부착 말뭉치 구축

(1) 사업 목표

○ 추론(inference)이 가능하도록, 사건을 중심으로 한 세밀한(fine-grained) ETS 정보가 부착된 20만 어절 규모의 한국어 평가 말뭉치 구축

○ 다문서, 다국어, 대화문과 같은 분산 텍스트 환경에서 추론이 가능하도록 함.

- 단일 문서 내에 표현된 ETS 정보의 상호 관련성 파악
- 다문서 및 다국어와 같은 분산 텍스트 환경에서 유관 사건에 대한

ETS 정보 파악

○ ETS 정보 부착 말뭉치는 어절 수보다는 문서를 기준으로 환산하여 결과물 구축 목표를 설정해야 함.

- 장르: 5개
- 문서: 장르별 100개
- 문장: 5장르*100문서*30문장 = 15,000 문장
- 어절: 15,000문장*13.3어절= 약 20만 어절

○ ETS 정보의 태그 세트 개발

○ ETS 정보의 자동 검색 방법론 연구

○ 태그 정보 부착을 위한 말뭉치 구축도구 개발

○ ETS 정보 부착 말뭉치 검색 및 활용도구 개발

(2) 세부 사업 내용

○ 국어정보화 1단계 사업 결과물(세종계획) 중 ETS 정보 부착 대상 문서 선별

○ 국어정보화 2단계 사업 내용 중 논항 정보 부착 말뭉치 및 화행 정보 부착 말뭉치 결과물 활용

○ 분산 텍스트 환경 평가에 적합한 말뭉치 추가 수집

○ ETS 정보 부착을 위한 태그 설계

○ 태그 정보 구축도구 개발

○ 말뭉치 구축도구를 이용한 ETS 정보 태깅 작업

○ 말뭉치 검색 및 통합적 활용도구 개발

2.2.1.3. 화행정보부착 말뭉치 구축

(1) 사업 내용 정의

○ 담화나 대화의 의미적 특성을 파악하는데 필수적으로 기여할 수 있는 화행정보 말뭉치를 구축하는 것임.

○ 말뭉치의 구축, 검증, 활용에 필요한 제반 도구를 개발함.

(2) 사업 목표

○ 향후 5년간 화행정보가 부착된 총 50만 어절 규모의 현대 한국어 말뭉치 구축.

○ 기존 구축된 말뭉치와의 연계성 등을 고려하여, 세종 구어 말뭉치 최대한 활용.

○ 범용성 높은 표준적 태그 설정 (ISO에서 논의중인 국제 표준을 최대한 준수).

○ 태그정보 부착을 위한 말뭉치 구축, 검증도구 개발

- 구문분석 말뭉치 구축 도구
- 화행정보 말뭉치 구축도구
- 화행정보 말뭉치 검증도구

○ 화행정보 부착 말뭉치 검색, 활용도구 개발

- 화행정보 말뭉치로부터 형태소, 구문, 논항 관련 정보 추출 도구

(3) 세부 사업

○ 화행정보 부착 말뭉치 주석 방식에 대한 국내외적 기준 조사 및 기준 확립

- 국제적 기준에 대한 포괄적 조사
- 국내 관련 학계의 용어 및 분류 기준 등 조사

○ 시험 말뭉치 구축

○ 태깅 도구 개발 및 운용 시험

- 태깅 검증 도구 개발 및 시험 구축된 말뭉치 검증

○ 수집 대상 자료 목록 선정

- 균형성 및 다양성 문제 검토
- 기존 세종 말뭉치(구어) 최대한 활용

○ 말뭉치 본격 구축

- 구문분석과 화행분석

○ 말뭉치 활용 방안 수립

- 활용 방향 정리 및 기초 활용 도구 개발
- 말뭉치 활용방법 홍보
- 사용자 의견 수렴 및 반영

2.2.1.4. 논조정보부착 말뭉치 구축

(1) 사업 내용

○ 본 사업의 주요 내용은 한국어 문서의 논조를 자동으로 파악하는데 사용하기 위한 한국어 논조분석 말뭉치를 구축하는 것임

○ 논조정보분석(Sentiment Analysis)은 최근 급증하는 블로그, 게시판, 웹 신문 등과 같은 온라인 매체상에 등장하는 특정 테마에 대한 사용자의 의견을 혹은 논조를 자동으로 분석하기 위한 연구분야임

○ 이를 위해 산업/학술 분야에서 논조분석에 대한 수요가 가장 큰 분야를 선정하고, 각 분야별로 대표적인 웹사이트, 게시판, 블로그 등에서 말뭉치를 수집함

○ 각 분야별로 수집된 말뭉치에 대해 주관적 문장과 객관적 문장을 구분한 후, 주관적 문장에 대해 긍정/부정/혼합 등의 논조태그를 부착함

(2) 사업 목표

○ 휴대폰, 자동차, 컴퓨터 등과 같이 사용자의 여론 혹은 버즈(Buzz)에 대해 정밀 분석을 요구하는 분야별로 논조정보가 부착된 총 1천만 어절 규모의 한국어 말뭉치 구축

○ 문장의 주관성/객관성 판별을 위해 각 문장에 대한 주관성/객관성 태그 부착

○ 주관적으로 태그된 문장에 대해 긍정/부정/혼합 등의 태그정보 부착

○ 태그정보 부착을 위한 말뭉치 구축도구 개발

○ 논조정보 부착 말뭉치 검색 및 활용도구 개발

(3) 세부 사업

○ 웹로봇을 통한 인터넷 사이트 문서 수집

○ 수집된 문서에 대한 정련작업

○ 논조정보 부착을 위한 논조태그 설계

○ 태그정보 구축도구 개발

○ 말뭉치 구축도구를 이용한 주관성/객관성 태깅작업

○ 말뭉치 구축도구를 이용한 주관적 문장에 대한 논조태그 부착

○ 말뭉치 검색 및 활용도구 개발

2.2.1.5. 다의어 의미부착 말뭉치 구축

(1) 사업 내용

○ 본 사업의 주요 내용은 다의어 수준의 어휘의미 정보를 부착한 한국어 말뭉치를 구축하는 것이다

○ 세종 1단계 사업을 통해 구축된 의미정보부착 말뭉치를 다의어 수준 정보부착 말뭉치로 확장함

○ 이를 위해 세종 1단계 사업의 의미체계와 표준국어대사전의 의미체계의 매핑을 시도함

○ 위 작업에 기반한 수정 표준국어대사전 의미체계에 기반하여 다의어 수준의 의미정보부착 말뭉치를 구축함

○ 다의어정보 부착 말뭉치 구축 및 활용을 위한 도구 개발

(2) 사업 목표

○ 표준국어 대사전 의미정보에 기반한 다의어 수준 의미정보가 부착된 1천 1백만 어절 규모의 한국어 의미정보부착 말뭉치 구축

○ 표준국어 대사전 의미체계와 세종 1단계 의미체계의 통합

○ 말뭉치 구축 및 활용도구 개발

(3) 세부 사업

○ 표준국어 대사전 의미체계와 세종 1단계 의미체계의 통합
○ 세종 1단계 의미정보부착 말뭉치 분석
○ 세종 1단계 의미정보부착 말뭉치에 다의어정보 부착
○ 태그정보 구축도구 개발
○ 말뭉치 검색 및 활용도구 개발

2.2.2. 호환성을 갖춘 한국어 어휘의미망 구축

2.2.2.1. 사업 목표

○ 호환성을 갖춘 한국어 범용 어휘의미망 (Interoperable KWordNet, 이하 I-KWordnet) 구현 (어휘의미 10만 개 내외)
  - 한국어 범용 어휘의미망 간 호환성
  - 국가 주도 언어자원(표준국어대사전, 세종전자사전)과의 호환성
  - 다국어 어휘의미망과의 연계성
  - 의미정보 부착 말뭉치와의 연계성
○ 어휘의미망 구축 및 관리 도구 개발
○ 어휘의미망 서비스 도구 개발

2.2.2.2. 세부 사업

(1) 한국어 범용 어휘의미망 간 호환성 확보
○ 호환/(연계/ 사상) 범위
  - Upper Ontology/ 개념망: 예) 명사 계층관계 1-4 단계
  - Main Body: 예) 명사 계층관계 4-10 단계
  - Domain Specific Terminology/ 전문분야 온톨로지: 의미 분야, 명사 계층관계 10단계 이하
○ 방식
  - 중심 어휘의미망 + 타 어휘의미망의 보완
  - 어휘의미의 크기(grain size) 가 유사한 어휘의미망 간의 사상
  - 호환 범위에 따라 상이한 방식 선택

(2) 국가 주도 전자사전과의 연계성 확보
○ 국가 주도 전자사전의 범위
  - 표준국어 대사전: 비교적 상세한 뜻풀이 말 포함하고, 특정 이론에 경도되지 않은 장점을 가지나, 자연언어 처리에 필요한 통사 및 의미 정보는 매우 간략함.
  - 세종 전자사전: 논항 정보와 선택제약 정보 등 자연언어처리에 필요한 다양한 언어정보 포함하고 있고 다의어 수준까지 구분된 의미가 기술되어 있으나, 뜻풀이 말이 아닌 의미부류 명으로 기술되어 있음
○ 전자사전과 통합 어휘의미망 간의 연계 방법
  - 어휘의미망의 각 노드 정의는 표준의 뜻풀이 말을 중심으로 연계함.
  - 세종의 의미 부류와 I-KWordnet의 upper ontology와의 사상을 시도하고, 세종에 풍부한 논항정보와 선택제약 정보를 I-KWordnet의 어휘의미에 연계함.

(3) 다국어 어휘의미망과의 연계성 확보
○ 범위
  - 다국어 연계성을 가진 어휘의미의 수: 7만 개 내외
  - 어휘의미의 알갱이 크기(grain size): 다국어 처리에 쓸 수 있을 정도로 세밀해야 함.
○ 방식
  - 다국어 연계성을 확보한 기개발 한국어 어휘의미망을 중심으로 연결함.

(4) 의미정보 부착 말뭉치와의 연계성 확보

○ 논항정보 부착 말뭉치

- 세종사전의 논항정보/선택제약정보가 연계된 어휘의미망을 기준으로 논항정보 부착 말뭉치를 개발함.

○ 화행정보 부착 말뭉치

- 용언 및 서술성 명사에 화행정보 연계

○ 개체/시간/장소 정보 부착 말뭉치

- 개체와 장소의 유형 분류에 어휘의미망의 계층 관계를 연계함.

○ 논조 정보 부착 말뭉치

- 어휘의미에 긍정/부정 가치를 부착함(SentiWordNet 참조).

○ 다의어 정보 부착 말뭉치

- 다의어 구분의 준거를 어휘의미망과 연계함.

2.2.3. 의미분석 자원 간의 연계/연동 시스템 구축

2.2.3.1. 사업 목표

○ 21세기 세종계획의 핵심 결과물인 세종 전자사전을 다양한 한국어 전산처리 과정에서 효과적으로 적극 활용하고 응용하기 위한 제반 기반을 구축

○ 표준국어대사전의 정보와 세종 전자사전의 정보 간의 연계 시스템을 구축하여 국가지원으로 구축된 언어자원의 효율적, 상호보완적 활용과 응용의 토대를 구축

○ 세종 말뭉치와 세종 전자사전 간의 연동 시스템을 구축하여 핵심 언어자원의 활용도와 효능을 극대화하는 환경 구축

○ 범용 한국어 어휘망 구축 사업의 기반 조성 및 지원을 위한 세종 전자사전의 의미부류 체계와 어휘의미관계 기술 체계의 확장 및 정교화

○ 의미정보 부착 말뭉치(다의정보 부착 말뭉치, 논항정보 부착 말뭉치 등) 구축 사업의 기반 조성 및 지원을 위한 세종 전자사전의 의미 및 논항정보 기술 체계의 확장 및 정교화

2.2.3.2. 사업 내용

○ 표준국어대사전 정보와 세종 전자사전 정보 간의 연계 시스템 구축

○ 세종 말뭉치와 세종 전자사전 간의 연계/연동 시스템 구축

○ 세종 전자사전의 의미부류 체계와 어휘의미관계 기술의 확장 및 정교화

○ 세종 전자사전의 논항정보 기술 체계의 확장 및 정교화 작업

○ 세종 전자사전 구축 도구 및 검증·검색 도구의 개선 및 통합 환경 구축

○ 세종 전자사전 기반 자연어 처리 시스템 모델 개발

2.2.3.3. 세부 사업

(1) 표준국어대사전과 세종 전자사전 간의 연계 시스템 구축

○ 표준국어대사전과 세종 전자사전의 표제어 매핑(동형어 정보 매핑)

○ 표준국어대사전과 세종 전자사전의 다의어 정보 매핑

○ 표준국어대사전과 세종 전자사전의 표제어의 의미별 정보 매핑

(2) 세종 전자사전과 세종 말뭉치 간의 연계/연동 시스템 구축

○ 세종 전자사전과 세종 말뭉치의 태그세트 통일

○ 세종 전자사전의 예문과 세종 말뭉치와의 연동 시스템 구축

○ 한국어 문서의 정보추출 및 문서분류 등 semantic web 구현의 기반 제공

(3) 세종 의미부류 체계의 확장 및 정교화

○ 세종 전자사전의 의미표상에 사용된 의미부류 체계의 정교화 및 통일화(명사의미부류 및 용언 의미부류)

○ 세종 전자사전 내 의미표상의 정밀성 및 체계성 제고

○ 세종 전자사전 내 모든 어휘에 대해 다의어 분할 층위에서 의미부류 부착 완료

○ 자연어 처리 과정 시 의미해석 및 정보추출 기능 제고

○ 한국어 어휘망 구축 및 반자동 온톨로지 구축의 기반 제공

○ 다의어 의미정보 부착 말뭉치 구축의 기반 제공 및 구축 사업 지원

(4) 세종 전자사전 내 어휘 의미 관계 기술 체계의 확장 및 정교화

○ 동의어 및 반의어 기술 체계 : 기존 기술 체계의 확장 및 개선을 통한 관련 정보 기술의 정교화 및 통일성 제고

○ 상위어 , 하위어 및 동위어 기술 체계 : 제1기 사업 시 고안되었으나 기술 유보된 상기 항목의 기술 체계의 구축을 완료하여 체언사전과 용언사전의 어휘의미 관계 정보에 기술을 완료

○ 한국어 어휘망 및 다국어 어휘망 구축의 기반 제공 및 구축사업 지원

(5) 세종 전자사전의 논항정보 기술 체계의 확장 및 정교화 작업

○ 의미역 기술 체계 및 통사 구문 정보 기술 체계의 확장 및 정교화

○ 논항선택제약 정보 기술 체계의 확장 및 개선을 통한 관련 정보의 정교화 및 체계화

○ 체언사전과 용언사전의 관련 정보의 유기성 및 통일성 제고

○ 한국어 구문분석 및 의미해석 시의 성능 및 활용성 제고

○ 논항정보 부착 말뭉치 구축의 기반 제공 및 구축 사업 지원

(6) 기구축 주요 하위 전자사전 개선 및 확장

○ 기구축 하위사전 중 중요도 및 활용도를 고려하여 선택된 소수 하위사전들에 대해 우선적이고 집중적인 개선· 확장 작업

○ 우선 작업 대상 하위사전 : 체언사전, 용언사전, 특수어사전, 고유명사사전

○ 각종 자연어 처리 과정에 필수적이고도 핵심적으로 활용되는 하위사전의 성능 및 활용성 제고 : 체언 사전 및 용언 사전

○ 자연어 처리 과정에 즉각적으로 활용되는 하위사전의 커버리지 확장 및 성능 향상을 통해 특수 목적의 전자 사전 구축을 위한 기반을 조성 : 특수어 사전 및 고유명사 사전

(7) 세종 전자사전 구축 및 검증·검색 도구의 개선 및 통합 환경 구축

○ 입력기 개선 및 통합 환경 구축

- 세종 전자사전 구축을 위해 개발된 하위사전별 입력기의 개선을 통한 통합 입력 환경 구축
- 제1단계 사업 시 개발된 통합입력기의 통합환경 개선과 정교화
- 특수목적 전자사전 등 추후 개발될 응용 전자사전의 입력 도구의 모델 제공

○ 상세 검색기 개선 및 통합 환경 구축

- 세종 전자사전 구축을 위해 개발된 하위사전별 상세 검색기 개선을 통한 통합 검색 환경 구축
- 체언 상세 검색기와 용언 상세 검색기의 통합 검색 환경 집중 제고
- 특수어 사전과 고유명사 사전 간의 통합 검색 환경 구축

○ 단순검색기와 상세 검색기 간의 통합 검색 환경 구축

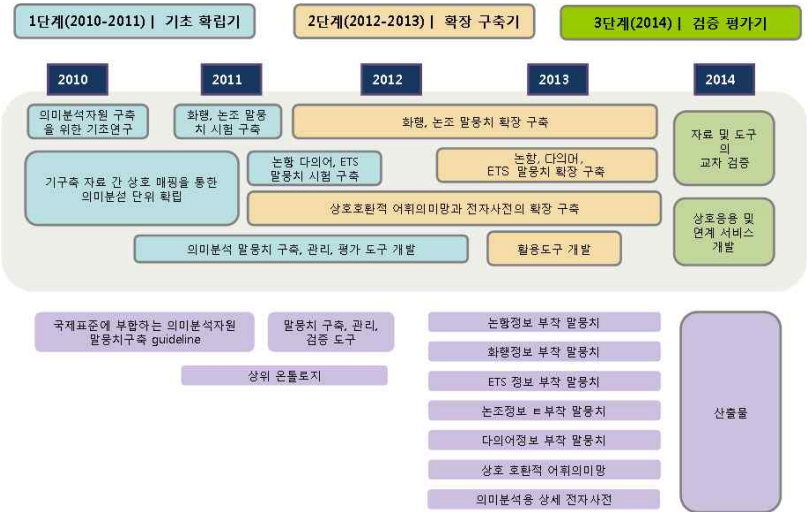
- 세종 전자사전 기반 각종 자연어 처리 시스템의 정보추출 및 활용의 알고리즘 제공

(8) 세종 전자사전 기반 자연어 처리 시스템 모델 개발

○ 제1기 사업 시 세종 전자사전의 성능 검증용으로 개발 시도되었던 각종 자연어 처리 시스템 모델의 개발을 완료

- 형태소 분석기
- 구문 분석기
- 의미 해석기
- 정보 검색 및 추출 시스템
- 기계번역 시스템
- 반자동 온톨로지 구축 시스템
- 한국어 기계 처리 및 다국어 기계 처리 시스템 개발의 기반 제공

2.3. 단계별 로드맵



3. 세계와 소통하는 한국어

3.1. 사업의 필요성

3.1.1 전문지식의 보편화와 자생적 융합을 위한 전문용어 표준화 체계 구축

○ 전문지식은 용어로 표현되며, 용어는 일상생활에서 자생적으로 쓰여지고 만들어질 때에 전 국민적 전문지식이 증가하여 사회발전과 국가지식관리가 가능해진다. 선진국일수록 일상용어와 어렵지 않은 전문용어가 매우 겹치고 있음.

○ 컴퓨터 “윈도”는 영어에서는 “window" 즉 ”창“이다. ”창으로 드려다 보는 컴퓨터 안의 모습“이라는 유사성으로 쉽게 접근할 수 있다. 그러나 ”윈도“라고 하면 주변 개념이 떠오르지 않아 창의적 발상이 중단됨.

○ 전문용어 표준화는 배타적 순우리말화보다는 현재 쓰는 언어에 기반하여 “조화”를 이루는 방향으로 진행되는 것을 ISO 704라는 전문용어 표준화 지침에서 추천하고 있음.

○ 전문용어를 쓰는 중요 정부기관, 언론의 용어와 더불어, 과학기술의 핵심인 기술표준원의 KS 용어, 특허청의 특허에 쓰이는 용어를 연계하는 정부차원의 범부처적 체계가 필요함.

○ 특히 교육에 초등학교 때부터 쓰여야 하므로, 표준화된 용어는 교육과학기술부의 교과서 편찬에 쓰이도록 해야 함.

3.1.2 국어 언어자원관리 표준 구축 및 생활 전문용어 구축

○ 공유되는 언어자원은 국제적 표준에 대응되어야 함. 기초적 언어자원을 국제적으로 쓰이도록 하여 국어에 대한 연구가 전 세계적으로 이루어지고, 또 국내외 서비스를 촉진하도록 함.

○ 언어자원의 국제표준화는 세 가지 의미가 있음.

- 국제적으로 표준화하려는 언어자원의 종류와 기술 수준에 동승할 수 있음.
- 국제표준화되어 있는 다른 언어와의 소통을 의미함. ISO/TC37는 개념중심의 다국어 표준화를 꾀하고 있음.
- 언어자원구축을 위한 지침 표준화에 따르는 인증된 자원을 만들 수 있음.

○ 언어자원표준화에 따라 국어 언어자원을 대응하도록 하고, 또 국어 언어자원에 따른 경험과 이론을 국제적으로 선도할 수 있어 국어정보화와 선진언어문화를 선도하여 미래지식정보사회와 산업을 이끌 수 있음.

○ 국민들이 금융, IT, CT, BT, NT 관련 용어, 방통 융합과 디지털 신기술 용어에 대한 이해가 너무 낮기 때문에, 이러한 생활 전문용어를 표준에 맞게 정비하여 대국민 서비스해야 할 필요가 있음.

○ 더 나아가 생활 전문용어를 자동으로 수집할 수 있는 도구의 개발이 요구됨.

3.1.3. 언어처리 기술 평가를 위한 표준화된 평가 세트 구축

○ 지식 정보 사회에서 정보의 양이 방대해짐에 따라서 대량의 정보에서 필요한 정보를 추출하는 정보검색 문제는 매우 중요하고 유용한 문제로 대두됨.

○ 인터넷 사회에서 정보검색은 가장 많이 사용되는 컴퓨터 기능중의 하나가 되어가고 있으며, 특히 국가 산업적인 측면에서 정보검색/추출/구축 기술은 매우 주요한 핵심 기술이 되고 있음.

○ 아울러 여러 시스템이 개발되는 과정에서, 개발된 정보검색 시스템의 평가, 시스템 간의 비교가 매우 중요한 문제가 되고 있음.

○ 그러나 이를 위해서는 평가 세트가 필요하나 이의 구축은 많은 시간과 인력을 요하는 힘든 문제임.

○ 선진국은 이미 이러한 문제를 오래전에 인식하고 국가적 차원에서 지원하고 있음.

○ 대표적인 예로 TREC 은 미국 정부 기관인 NIST 의 조직 및 지원으로 열리는 정보검색 관련 학술대회로서 평가 세트를 제공하여 시스템 간의 비교 평가를 통한 기술 발전을 유도하고 있으며, 일본의 NTCIR도 그러함.

○ 국내에서는 KISTI에서 수년 전에 HANTEC 평가 세트를 개발하였으나 지속적인 유지 보수 관리 및 개선이 이루어지지 못하고 있음.

○ 문서분류는 대량의 문서를 여러 범주로 분류하여 줌으로써 관심 범주의 문서만을 다룰 수 있도록 하는 주요한 기술이며, 문서분류 시스템 기술의 발전을 위해서는 정보검색과 마찬가지로 평가 세트가 필요함.

○ 선진국은 이를 위해 많은 투자를 통해 이미 많은 수의 문서 분류 평가 세트를 구축하여 활용하고 있으나, 국내에는 이에 대한 준비가 매우 부족함.

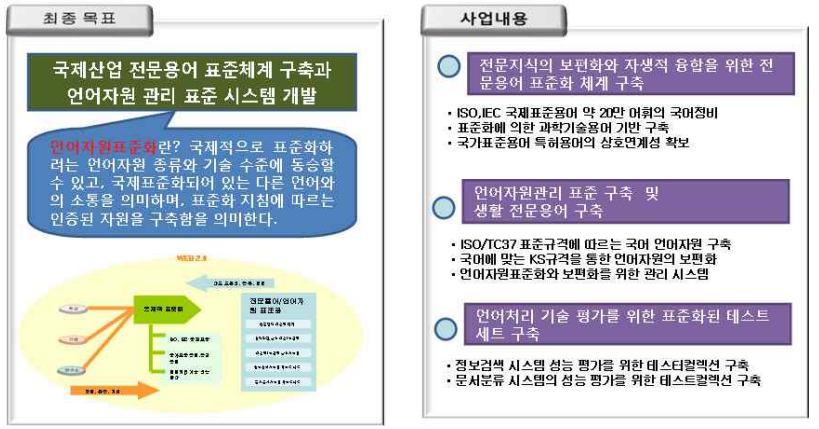
○ 이를 위해서 한국어의 세계화 및 그 효율성 제고를 위해서는 한국어 평가 세트를 준비하여 활용하는 것이 필수적임.

3.2. 사업 내용

○ 국제적인 국어 지식 정보 기반으로써 국제산업 전문용어 표준체계를 구축하고, 용어 표준 서비스 기반 및 언어자원 표준화 기반 서비스 시스템을 개발함.

○ 또한 생활 전문용어의 정비와 대국민 서비스를 목표로 함.

○ 정보검색 및 문서분류 시스템들의 성능 평가 및 비교를 가능하게 하여 관련기술의 발전에 필수적인 평가 세트를 개발하는 것을 목표로 함.



3.2.1. 전문지식의 보편화와 자생적 융합을 위한 전문용어 표준화 체계 구축

3.2.1.1. 사업의 목표

○ ISO, IEC 국제표준용어 약 20만 영어 용어의 국어 정비 및 표준화에 의한 과학기술용어 기반 구축

- 영어와 국어의 난이도에 따른 대응수준 체계화
- 개념 중심의 일관성 있는 국어 용어 체계화
- 용어표준규격의 전문용어에 대하여 온톨로지화

○ 국가표준용어, 특허용어의 상호연계성 확보

- KS 표준화 용어가 특허청의 특허제목과 요약에서 상호연계가 되도록 함

3.2.1.2. 세부 사업 내용

○ ISO의 표준용어는 2005-2006년도에 한중일 표준기관의 협조하에 데이터베이스가 만들어져 ISO 행정부에서 Conceptual DB라는 이름으로 2009년 공개 예정되고 있음. 이에 대한 한국어 대응을 일차목표로 함.

○ ISO의 용어는 과학기술, 경영 경제, 언어, 약의학, 문헌정보, 지식관리, 물류, 소방, 조선, 전기전자 등에 이르는 다양한 산업화에 필요한 분야를 망라하고 있음.

○ ISO 용어표준이라 함은 용어규격과 일반규격의 용어정의부를 의미함. 용어규격은 그 분야에서 필요한 필수적 용어를 개념체계 아래에 정의하고 있음. 일반규격의 용어정의부는 용어규격 및 다른 일반규격에서 정의하지 않은 것을 다른 규격용어를 참조하여 정의하고 있음.

구분	세부업무	연구개발내용	연구개발범위
개발	ISO용어, 특허요약문, 학단연용어, 국립국어원 용어, 사전 등이 통합DB화	현존하는 용어자원을 수집하여 일관성 체계구축	100만 용어 추산
연구	개념중심의 재분류	개념표현, 조어와 생성 규칙화 및 적용실험	같거나 비슷한 용어 군집화
개발	영어와 국어의 일상용어 대응체계	영어-국어 어휘수준의 대응화	ISO대상용어의 구성요소 어휘에 따름
개발	한국어-영어 대비 대응구축	개념과 어휘수준에 따른 영-한 용어 대응	어휘 매핑 및 사람에 의한 검증 작업
연구	용어와 주변어휘 및 동사와의 조화 측정	용어의 쓰임새를 위하여 명사-동사의 원활한 활용 개발	ISO용어규격 정의문에 국한함
개발	ISO, 특허 용어와의 대응체계화	특허제목과 요약의 영한대역본을 중심으로 용어표준정비 체계화	ISO용어에 따라 적용범위 결정

3.2.2 국어 언어자원관리 표준 구축 및 생활 전문용어 구축

3.2.2.1. 사업의 목표

○ ISO/TC37 표준규격에 따른 국어 언어자원에 대한 적용 및 국어에 맞는 KS규격을 만들어 언어자원의 보편화.

○ 국어 언어자원표준 규격에 맞는 생활 전문용어 구축과 이의 대국민 서비스

3.2.2.2. 세부 사업 내용

○ ISO/TC37은 전문용어 개발 지침, 인증, 실제 참고문헌과의 전자표준, 사전형식 및 편찬 지침, 언어자원 형식표준과 각 언어계층인 형태-구문, 의미, 담화, 시간, 공간, 다국어화, 사전형식에 대한 표준이 있음. 이에 대한 국어 언어자원 및 언어자원이 내재할 수 있는 콘텐츠 자원에 대한 표준화 체계를 구축함

○ ISO/TC37의 언어별 표준 중의 하나인 “단어분절원칙”에 대한 표준 및 데이터범주 표준 등은 각 언어에 대한 적용이 필수적이므로 이에 대응 표준 및 언어자원 개발이 표준화되도록 함

○ 국민들이 자주 접하는 경제, IT, BT, NT, 방송통신 등의 분야에 대한 최신의 생활 전문용어를 5만 어휘 수준으로 구축

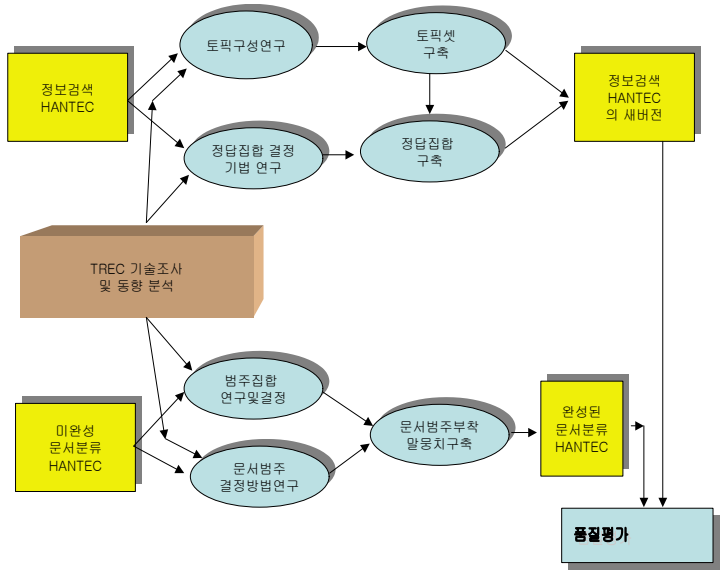
구분	세부업무	연구개발내용	연구개발범위
연구/개발	국제표준에 대한 대응체계화	언어자원개발, 인증, 형식표준에 대한 국어대응체계화. 국제 표준의 적용	현재 해당 국제표준 범위 내. 관련 시스템 및 프로그램 개발
개발	실용 전문용어 구축	표준 적용	실용 전문용어 구축 및 자동추출프로그램 개발
개발	정부 및 민간 교육, 지침 체계구축	주석 (어노테이션), 전문가 양성 및 교육체계화	언어자원에 따른 보급화 교육 체계

3.2.3. 언어처리 기술 평가를 위한 표준화된 평가 세트 구축

3.2.3.1. 사업의 목표

○ 정보검색 시스템의 성능 평가 및 비교를 위한 평가 세트를 구축함.
- HANTEC 정보검색 평가 세트의 평가
- 보유 평가 세트의 정제를 위한 전략 및 기술 연구
- HANTEC 정보검색 평가 세트의 개선 및 정제

○ 문서 분류 시스템의 성능 평가 및 비교를 위한 평가 세트 구축
- 문서집합 구축
- 범주 집합 설정
- 범주 부착 말뭉치 구축



○ 본 연구의 목표는 정보검색 및 문서분류 시스템들의 성능 평가 및 비교를 가능하게 하여 관련기술의 발전에 필수적인 평가 세트를 개발하는 것을 목표로 함.

구분	세부업무	연구개발목표
개발	기존의 HANTEC 정보검색 평가 세트에 대한 개선 및 정제	정보검색 평가 세트구축
개발	HANTEC 문서분류 말뭉치에 대한 범주부착 말뭉치 구축	문서분류 평가 세트구축

3.2.3.2. 세부 사업 내용

○ 지난 수년 동안 중단된 한글 정보검색 평가 세트(HANTEC)의 개선 및 정제를 통하여 국제적 규모의 한국어 검색 비교평가 기반 구축 **- 매년 10만 건 수준**

○ 다범주 문서분류(Text Categorization) 평가 세트 구축을 통한 한국어 문서범주기의 기본 성능 평가 기반 구축 **- 매년 5만 건 수준**

구분	세부업무	연구개발내용	연구개발범위
연구	토픽의 구성 각 필드의 내용에 대하여 연구함	HANTEC 토픽집합 검토 및 정제	50 개의 토픽
개발	질의에 대한 정답문서를 결정함	HANTEC 정답문서집합 구축	질의에 해당하는 정답문서집합 구축
개발	HANTEC 정보검색 문서의 개선 및 정제	HANTEC 정보검색 문서집합의 갱신 및 재구축	매년 10만개의 문서 개선 및 정제
연구	문서분류 시스템의 성능 평가를 위한 실험세트의 구축을 가능케 하는 기술을 연구함	문서분류 평가 세트 구축 방법론 연구	구축기술
연구	주어진 문서집합에 대한 범주집합의 granularity 결정 및 범주집합을 결정함	문서분류 평가 세트의 범주집합 결정	10~20개의 범주집합 결정
개발	대량의 문서에 대한 범주를 결정하여 부착하는 작업을 수행함	문서분류 평가 세트의 정답 범주 레이블링 부착 작업	매년 5만문서 범주부착 문서집합 구축

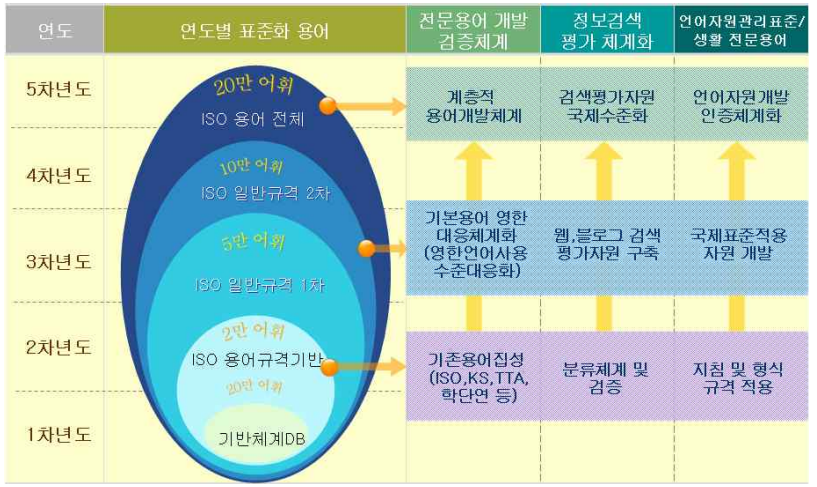
3.3. 단계별 로드맵

○ 전문용어, 정보검색 평가체계, 언어자원의 국제표준을 서로 연동하여 연구개발하도록 단계화함.

○ 초기 전문용어는 현존하는 용어를 집성하는 단계가 중요함. 그 양은 추산하여 적어도 20만이라고 가정함.

○ 정보검색 평가체계는 각 개발하는 전문용어 수준과 언어자원 국제표준체계와 연동하여 개발하는 것을 원칙으로 함.

○ 초기 전문용어 정비 및 표준화는 매우 기본적인 용어대응, 언어수준 대응 등을 집중 연구개발하여 양보다는 질의 토대를 마련하도록 유도함.



V. 기대 효과 및 활용방안

1. 기대 효과

1.1 사회적/교육적 측면

○ 과거는 단순히 발표문 형식으로 국가 언어 환경에 영향을 줄 수 없었지만, 이를 통해서 국어 표준화 및 바른 우리말 사용 등을 좀 더 용이하게 독려할 수 있음.

○ 포털에서 이 인터페이스를 도입함으로써 블로그 및 카페, 덧글을 쓸 때 실시간으로 자신의 말을 철자검색할 수도 있으며, 순화용어에 대한 교육도 받을 수 있을 것임.

○ 발생하고 있는 언어현상에 대한 통합적인 분석이 가능하게 되어 잘못 사용되고 있는 한국어 어휘에 대한 동향 조사를 통해 바른 한국어 보급을 위한 방향을 신속하고 빠르게 설정할 수 있음.

○ 이를 통하여 한국어의 올바른 발전을 꾀할 수 있음.

○ 국어정보화 1단계 사업이 시작될 때, 원시 말뭉치와 형태소 분석 말뭉치가 연구실/연구소나 개인 연구자 차원에서 산발적으로나마 개발한 경험이 비교적 짧은 기간 내에 21세기 세종계획 연구성과를 내는 데 중요한 밑거름이 되었듯이, 현재 의미정보에 관한 연구와 말뭉치 역시 연구실/연구소, 개인 연구자 차원에서 연구 역량이 충분히 축적되어 있으며, 국제적 협동연구 기반도 다져 놓았음. 이제 이를 결집하여 국가 차원의 대규모 정교한 의미관련 정보 말뭉치 및 어휘의미망을 구축하여 소규모 연구실/연구소, 개인 연구자 및 언어관련 기업에 제공한다면, 미래 지식기반 사회에서 요구하는 다양한 의미처리 및 지식처리 관련 기술 및 서비스를 빠르고 다양하게 국민에게 제공할 수 있음.

○ 국외에서 의미정보 연구와 말뭉치 구축은 2-30년에 걸쳐 동일한 개발자들이 꾸준하게 진행해 옴. 한국어에 대한 의미처리 연구와 말뭉치 구축은 15-20년 정도 늦게 시작하였으나, 본 과제가 진행된다면 매우 강력해진 하드웨어 및 소프트웨어 환경과 더불어 현재 연구실/연구자 별로 추진해 온 국제적 협동 연구를 통해 선발 연구 그룹의 경험과 시행착오를 타산지석으로 삼아 비교적 단기간 내에 국외 연구 그룹과의 기술 격차

를 줄여 미래 지식기반 사회를 선도할 수 있음.

○ 지금까지 국어 학습 시스템이 학습자에게 단순히 학습자료를 보여주고, 철자와 형태 정보 단계까지만 자동으로 교정하는 초급 학습 시스템 개발에 머물렀으나, 어휘의미망을 이용한 연관어 학습, 논항 정보 및 다의어 정보를 이용한 통사 교정 및 정교한 어휘 제시 등 고급 한국어 학습 시스템 개발이 가능함.

○ 전문용어는 지식을 담는 그릇으로 이를 표준 정비하여 외국어와 동등한 한국어 용어가 만들어진다면, 한국어 일상어휘가 더 늘어나고 국민의 지식과 생활의식이 향상될 것임.

○ 급증하는 생활 전문용어를 정비하여 제공함으로써 국민들의 이해도를 증진시키고, 이를 통해 국민의 지식 역량이 강화될 것임.

1.2. 기술적 측면

○ 현재 국내 포털 사업자들은 각자 독자적으로 언어자원을 축적하고 언어처리 기술을 개발하고 있음.

- 상대적으로 해외 기업들은 국내 기술을 구매하여 단기간에 유사한 품질의 서비스를 출시하고 있음.
- 국내 업체 간에 중복 투자로 인한 낭비를 줄이고 오히려 이를 보다 건설적인 추가의 언어자원을 구축하는데 투자할 수 있다면 그 효과가 더욱 극대화될 수 있을 것임.

○ 이를 위해서는 우선 공통으로 활용될 수 있는 언어자원을 구축하여 공유해 주어야 함.

- 주로 색인 기술과 이에 필요한 전자사전, 말뭉치 등이 가장 기초적인 언어자원이 됨.
- 각 업체들이 산발적으로 행하고 있는 품질 테스트용 평가 세트를 공신력 있는 기관에서 제공함으로써 중복 투자를 줄일 뿐만 아니라 국내 업체들의 검색 서비스 품질을 향상시키는 촉매제 역할을 할 수 있을 것임.

○ 학계와 기업에서 많은 언어도구 및 요소기술이 개발되고 있으나 이를 위한 공인된 평가 시스템이 없는 관계로 언어도구나 요소기술에 대한

평가가 제대로 이루어지지 않음.

- 평가가 이루어지지 않기 때문에 개발된 각 시스템에 대한 분석 및 개선이 어려움.
- 평가시스템을 만드는 것은 많은 자료 수집과 이에 대한 정답 Set 구축, 이를 이용한 평가 도구 개발 등의 시간 및 인력이 많이 드는 작업으로 학계나 기업이 단독으로 수행하기 어려움.
- 그러므로 언어도구 및 요소기술 평가 시스템 구축은 학계나 기업 등 기술을 개발하는 주체 모두에게 도움을 줌.
- 아울러 다음과 같은 이점을 기대할 수 있음.
  - 웹 정보검색 및 일반 정보검색 시스템 성능평가 및 비교를 가능케 함.
  - 이에 따른 정보검색 시스템 기술을 발전이 기대됨
  - 정보검색/문서분류 소프트웨어 상품화
  - 기존 상품화된 시스템 간의 품질 비교 및 경쟁 강화
  - 기술개발 단계에서 시스템의 측정, 시험을 통하여 고품질 제품 개발 유도
  - 정보검색/문서분류 소프트웨어 제품 개발 환경 향상

○ 기초언어분석기는 다양한 요소기술 과 서비스를 개발하기 위해 가장 기초가 되는 기술임.

- 그러나 현재까지는 기초언어분석기가 없거나 공유되지 않아 요소기술과 서비스를 개발하기 위해서는 기초 언어분석기까지 개발해야 했음.
- 따라서 기초 언어분석기가 오픈 프로젝트화 하여 일반에게 공개된다면 새로운 요소기술과 서비스를 개발하는 학계나 기업의 부담을 감소시켜 더 많은 요소기술과 서비스를 개발할 수 있음.
- 오픈프로젝트의 특성상 집단지성을 활용하여 다양한 기술 개선 아이디어 및 신기술 등이 기초언어분석기에 적용되어 공개되기 때문에 기초언어분석기 기술을 선도하게 됨.

○ 국가 기반 언어정보를 통해서 국가의 전체적인 언어자원이 통합관리되고, 더욱이 언어 현상에 대해서 표준화된 전산환경 관점의 대응이 가능하게 됨.

○ 초기 정보화 단계에서는 디지털화된 자료의 '양'이 중요했으나, 현재 유

용한 정보와 쓰레기 정보가 뒤범벅이 되어 홍수처럼 쏟아지는 현재에는 자료의 '정확성'과 '질'이 중요하며, 자료의 양이 기하급수적으로 폭증하게 될 미래에는 개인이 필요로 하는 정보만을 선별하는 맞춤형 정보 서비스가 각광을 받게 될 것임. 이를 위해서는 문장과 문서의 의미, 추론, 사용자 의도 파악에 관련된 기술이 필수적임.

○ 각종 의미정보 말뭉치와 어휘의미망 간의 교차 응용하여 새로운 언어자원과 활용 시스템을 개발할 수 있음. 예를 들어 논조 정보 말뭉치의 결과를 어휘의미망과 연결하여 Korean SentiWordNet을 개발할 수 있으며, 화행 정보가 부착된 대화 말뭉치와 ETS정보 말뭉치를 이용하면 다양한 대화형 MMI(Man-Machine Interaction) 시스템 개발이 가능함.

○ 정보검색 산업, 인터넷 산업, 특허 및 표준과 같은 국가 지식체계의 인프라 구축이 가능함.

○ 차세대 인터넷 사업이 더 빠른 인터넷에 그쳐서는 안 되고, 그 인터넷에 실릴 지식으로서의 언어자원 표준화와 개발이 가능함.

1.3. 산업적 측면

○ 포털 사이트와 공조하여 원시데이터(Raw Data)인 실생활에서 사용되는 블로그 포스트, 뉴스 콘텐츠, 질의어 등을 기초로 언어자원을 구축함으로써 실용적으로 활용될 수 있는 수준의 언어 자원을 확보하고 지속적으로 확장하는 체계를 마련함으로써, 단순히 학계뿐만 아니라 연구소, 기업체 등 산업 전반에 걸쳐 활용될 것임.

○ 블로그, 카페, 댓글 등은 인터넷이 대중화된 현재 일반인이 언어를 가장 많이 이용하는 분야일 뿐만 아니라 포털이나 기업 등에서 가장 정보처리가 필요한 분야임.

- 그동안 국가과제 등에서는 이런 현실적인 요구들이 간과되어 개별 기업, 특히 이런 콘텐츠를 보유한 포털 등에 의해 이 분야 언어자원이 구축되어 왔음.
- 따라서 각 기업들은 같은 언어 자원을 구축하는데 중복 투자를 하게 되는 문제가 있을 뿐만 아니라 작성된 자원을 독점하게 되어 언어 자원 이용의 효율성을 저해하고 있음.

- 본 사업에서의 실용언어 자원 구축은 기업들에 자원 구축에 대한 중복 투자를 줄이고 요소기술 및 서비스에 집중하게 하여 개발 속도 개선 및 양질의 서비스 개발에 도움을 줌.
- 실용언어 자원은 이미 포털을 포함한 인터넷 서비스, KMS, CMS, VOC 등의 다양한 지식기반 서비스 등에서 활용하기 위한 많은 연구 및 기술개발이 이루어지고 있고 대국민을 대상으로 한 많은 상용화 서비스가 현재도 제공되고 있으므로, 이들을 위한 실용언어 자원 구축은 국어자원의 활용도를 가장 높이는 방법임.
  - 인터넷 환경에서 국어 사용자들이 실생활에서 실제로 사용하는 언어와 공식적인 문서 등에서 사용되는 문어체를 비교-연구하는 데 중요한 자료로 사용됨.
  - 실생활에서 발생 빈도가 높은 철자 오류, 띄어쓰기 오류, 문법 오류 자료를 이용하여 우리말 사용 실태와 변화를 정확하게 측정하고 분석하는 데 유용한 정보를 제공함.
  - 블로그, 카페, 댓글 등 국어 사용자들이 직관적으로 사용하는 언어 현상을 정확하게 파악하는데 기여함.
  - 기구축된 말뭉치를 편리하고 효율적으로 활용하는 데 공통적으로 필요한 노력을 절감하는 효과가 있음.
  - 언어처리 분야의 연구 개발자 및 관련 연구자들이 필요로 하는 국어의 언어 현상을 쉽게 활용할 수 있도록 활용 가치가 높은 정보들을 추출하여 사용자의 요구 사항에 적합한 형태로 가공하여 제공함.
  - 실용언어 자원은 구축된 국어자원의 활용도를 높임
  - 정보처리를 위한 기초 인프라 제공으로 국어자원의 활용 증대
  - 국어자원을 활용한 양질의 기술 개발 지원
  - 국어자원 활용을 위한 기술인력 인프라 증대

○ 본 과제에서 결과물로 도출된 각종 의미정보 부착 말뭉치와 어휘의미망 및 세종전자 사전은 의미정보 처리용 분석/생성 시스템 개발을 위한 기계학습 데이터로 활용 할 수 있음. 특히 의미모호성 해소 모듈 개발을 원하는 업체 및 연구소에 유상으로 지원하여 향후 말뭉치 관리비용 충당할 수 있음.

○ 의미관련 정보 언어자원 개발은 언어 및 지식 처리의 초기 과제이자 궁극적인 목표 중의 하나인 기계번역 및 자동통역 시스템의 품질을 향상

함.

○ 언어처리와 지식처리 관련 기술의 응용과 발전 방향은 형태 기반에서 의미기반으로 변화하고 있으므로, 본 세부과제에서 구축하려는 다양한 의미정보 부착 말뭉치와 호환성이 있는 어휘의미망의 개발 및 세종전자사전의 완성은 관련 학계와 산업계가 요소기술, 분석/생성 도구, 서비스를 발전시키는 데 매우 중요한 기초 인프라를 제공하게 될 것임.

○ 정보검색 산업, 인터넷 산업, 특허 및 표준과 같은 국가 지식체계 구축이 기대된다.

○ 언어자원표준화에 따라 국어 언어자원을 대응하도록 하고, 또 국어 언어자원에 따른 경험과 이론을 국제적으로 선도할 수 있어 국어정보화와 선진언어문화를 선도하여 미래지식정보사회와 산업을 이끌 수 있음.

2. 활용 방안

2.1. 지식사회를 선도하는 한국어

(1) 정보 검색 서비스

○ 국내 포털 서비스는 유사한 검색 서비스로서 상호 경쟁하고 있음.

○ 그 원인은 새로운 검색 모델을 개발하기에는 많은 전문 인력과 시간을 요하고 또 기존의 모델을 바꿀 경우 사업적 위험도 존재하기 때문임.

○ 이러한 이유로 인해 새로운 검색 모델에 대한 시도가 이루어지지 않게 되고, 이로 인해 해외에 비해 한국의 검색 서비스 수준이 점차 경쟁력을 잃게 되어 장기적으로 부정적인 결과에 도달할 위험이 높아지게 됨.

○ 따라서, 시맨틱웹 검색 등의 보다 진화된 검색 모델에 대한 핵심 모델을 수립하고 공개하여 이러한 분야에 대한 연구/개발을 촉진할 필요가 있음.

○ 분야별 전문 검색과 같은 새로운 검색 모델을 지원하기 위해 필요한 자동 분류 기술을 제공함으로써 국내 검색 서비스의 활성화와 해외 사업자 대비 차별화 우위를 더욱 공고히 하는 효과를 기대할 수 있음.

○ 이러한 분야에 본 과제를 통해 제공되는 언어자원들이 기초 지식으로서

중요한 역할을 가짐.

(2) 콘텐츠 중복 제거 및 필터링

○ UCC의 증가로 인해 그 요구가 증가하고 있는 비정형 문서의 중복 인식 및 제거 기술, 자동 분류 기술 등을 제공함으로써 이를 통한 불필요한 정보가공 및 모니터링 비용을 절감하는 효과를 거둘 수 있고, 고객들은 보다 정확하고 깨끗한 서비스를 제공받을 수 있게 될 것임.

○ 최근 블로그 서비스의 대중화로 인해 블로그 포스트가 중요한 정보원의 역할을 하고 있는데, 상당수 블로그 포스트가 펌질을 통해 배포된 것이어서 검색시 중복이 발생하게 됨.

○ <그림 9>에서와 같이 15일간 수집한 블로그 포스트 중 9.28%가 중복이며, 과거 포스트와의 중복은 30% 정도로 예측됨.

○ 보다 높은 품질의 검색 결과를 제공하기 위해서는 이러한 중복된 포스트를 클러스터링하는 기술이 필요하며 이를 위해서는 다양한 분야의 용어가 기초 자료가 됨.

<그림 9> 국내 블로그 포스트의 중복 비율
(Search Day 2008 Spring, 온네트 박수정)

○ 또한 최근 블로그를 통한 바이럴마케팅이 확대되면서 스팸성 블로그 포스트도 양산되고 있는 추세임.

○ 스팸메일을 필터링하는 것과 유사하게 스팸 포스트에 대한 필터링도 중요한 기술로서 연구/개발되고 있으며 이 부분에도 실용 언어 자원이 중요한 역할을 하게 됨.

(3) 문맥의존형 광고

○ 최근 키워드 검색 광고의 차세대 모델로서 문맥의존형 광고와 관련한 기술에 대한 필요성이 증가하고 있음.

○ 오버추어/구글 등에서만 제공되고 있는 문맥의존형 광고는 한국어 처리 기술의 한계로 인해 해외 시장에 비해 국내 시장에서는 큰 성장을 하지 못하고 있음.

○ 이에 대해 국내 업체들이 새로운 시도들을 하고 있으나 언어자원 및 기술력의 한계로 가시적인 효과를 내지 못하고 있는 실정임.

○ 문맥 인식 기술과 이를 응용한 광고 플랫폼의 개발을 통해 지속적으로 증가하고 있는 검색 광고 시장을 국내 업체들이 확보할 수 있고, 나아가 해외 시장으로 진출할 수 있을 것임.

(4) 자원 공유 환경 제공

○ 앞서 기술한 언어자원 및 인프라, 원천 기술 등은 국내업체에 자유롭게 공유되고 활용될 수 있어야 할 것임.

○ 또한 국내의 다양한 전문가들이 지식을 공유하고 개선할 수 있는 환경을 제공함으로써 자생력을 가지고 발전할 수 있을 것임.

○ 언어자원의 공유뿐만 아니라 언어처리 기초 모듈들을 오픈 소스화하여 전문가들이 자발적으로 개선하고 활용할 수 있는 환경을 조성해 주어야 할 것임.

○ 이를 통하여 본 과제를 통해 축적된 자원들이 보다 넓은 분야에서 각 분야에 맞도록 커스터마이징되어 활용될 것임.

2.2. 미래를 준비하는 한국어

(1) 논항정보부착 말뭉치

○ 한국어의 논항과 문법기능, 또는 논항과 문법표지 사이의 체계적인 연관성을 파악하기 위한 데이터로 활용

○ 의미에 기반을 둔 한국어 어휘망 구축에도 중요한 기여할 수 있음.

(2) 개체·시간·공간(ETS) 정보 부착 말뭉치

○ 지식처리의 발전 방향인 자연언어처리 기반 추론에 대한 연구/개발 초석 제공

○ 국제적 협력 연구 체재의 기반이 마련되어 있으므로, 2단계 사업이 완료될 시점에는 이론과 실제에서 이 분야에서 국제적으로 최고 수준의 원천 기술력을 확보할 수 있음.

○ 검색 분야에서 현재 기술인 문자열 검색과 온톨로지를 이용한 연관어 확장 검색의 한계를 뛰어넘을 수 있음.

○ ETS 정보 자동 부착 시스템의 개발은 언어학의 기존 연구결과와 전산학의 기술력을 긴밀하게 접목해야 가능한 분야로서, 양 영역의 연구 활성화에 시너지 효과를 도모함.

(3) 화행정보부착 말뭉치

○ 상업적 활용도가 높은 질의응답 시스템 개발에 기초자료로 활용될 수 있음.

(4) 논조정보부착 말뭉치

○ 최근 급격히 수요가 증가하는 논조정보 자동분석분야에 대한 연구/개발 초석 제공

○ 논조정보부착 말뭉치에 기반한 기계학습방식의 적용 등을 통한 논조자동분석 시스템의 개발

○ 추가적으로 말뭉치로부터 논조 관련 어휘사전을 구축

○ 향후 한국어 감정정보 워드넷의 개발에도 이용

○ 언어학과 전산학 분야의 접목을 통한 새로운 연구분야에 활용

(5) 다의어 의미부착 말뭉치

○ 기계번역, 정보검색, 대화시스템 등과 같은 의미처리가 필요한 어플리케이션의 개발을 위해 반드시 필요

○ 다의어 정보부착 말뭉치에 기반하여 명사 뿐만 아니라 동사와 같은 용언의 의미모호성도 해소하여, 기계번역 시 대역어선택 문제 등과 같은 성능향상에 획기적인 영향을 미치는 분야에 기여할 것으로 기대

○ 향후 명사어휘망 뿐만 아니라 형용사, 동사 등과 같은 다른 품사의 어휘망 구축에도 응용될 수 있을 것임

(6) 호환성을 갖춘 한국어 어휘의미망

○ 구축하려는 I-KWordNet의 크기가 중형 사전에 해당하는 것으로 실제 자연언어처리 시스템의 다양한 분야에 활용 가치가 높음.

○ 어휘의미의 알갱이 크기가 충분히 세분화되어 다른 언어의 어휘의미와의 등가성을 설정하는 데 비교적 용이하며, 다국어 연계성을 확보하는 데 유리함.

○ I-KWordNet은 일반목적의 범용적 지식표상체계를 지향하므로, 인간의 상식이나 일반적 배경지식을 추론하는 데 적합함.

○ 지금까지 각양각색의 기준과 방법론에 따라 개발되었거나 개발될 전자사전, 말뭉치, 어휘의미망의 상호호환성과 연계성은 한국어정보처리에 대용량 고품질의 기초자원을 제공해 줌.

2.3. 세계와 소통하는 준비하는 한국어

(1) 전문용어 표준화 체계 구축

○ 학술 전문용어는 각 학문 분야의 지식을 담는 그릇으로 이를 표준 정비하여 외국어와 동등한 한국어 용어가 만들어진다면, 한국어 일상어휘가 더 늘어나고 국민의 지식과 생활의식이 향상될 것임.

○ 정보검색 산업, 인터넷 산업, 특허 및 표준과 같은 국가 지식체계에 활용함.

(2) 언어자원관리 표준 및 생활 전문용어 구축

○ 언어자원 국제표준은 전문가 그룹을 형성화하기 위한 교육체계, 연구개발의 보편화와 보급을 전제로 함.

○ 차세대 인터넷 사업이 더 빠른 인터넷에 그쳐서는 안 되고, 그 인터넷에 실릴 지식으로서의 언어자원 표준화와 개발에 활용됨.

○ 특정 논쟁(이슈)에 따라 다양한 매체를 통해 쏟아져 나오는 생활 전문용어에 대한 정확한 개념을 제공함으로써 대 국민 서비스에 활용됨.

(3) 언어처리 기술 평가를 위한 표준화된 평가 세트 구축

○ 표준화된 평가 세트를 이용하여 기 개발된 언어처리 지원 프로그램의 성능 평가 및 산업계에서 개발하는 각종 프로그램의 성능을 객관적으로 평가할 수 있음.

○ 국어정보화 2단계 사업의 진행 단계에서 공모전을 통해 우수 연구 개발팀 선정할 경우 활용함.

빈 페이지 임

VI. 추진체계

1. 추진전략

1.1. 사업 절차

○ 본 사업의 총괄은 국립국어원에서 사업관리팀(task force)을 구성하며, 각 사업 및 과제의 선정, 평가를 위해서 국어심의회에 별도의 사업심의회(국립국어원 담당 포함 산학연 전문가를 일정 비율로 구성)를 구성하는 것을 원칙으로 함.

○ 본 사업의 3개 사업 분야는 서로가 연계하는 부분이 많은 만큼 이의 원활한 소통(의견, 시스템 교류 등)을 위하여 본 사업의 추진 위원회(사업에 참여하지 않은 산학연 전문가 및 단위 과제 책임자)를 두며, 추진 위원회에서는 사업의 방향, 표준화(de facto 등) 문제점 해결 방안, 사업내용의 조정 등을 수행함.

○ 총 사업 수행 기간은 5년으로 하고 필요에 따라 단계(2-2-1년)을 구분함.

- 1단계(2010~2011, 2년) **기반 확립단계** : 각 세부 사업에 요구사항, 결과물 구축을 위한 정확한 지침을 준비함. 공모대회를 통해 최우수 언어처리 기반기술 보유팀을 선정함.
- 2단계(2012~2013, 2년) **확장 구축단계** : 세부 사업 진행
- 3단계(2014, 1년) **검증 및 평가단계** : 세부 사업 결과물을 대국민서비스, 상용화, 산업화를 준비하기 위한 단계

○ 본 사업의 RFP 작성을 위해서 사업 시작 전에 각 3개 사업 분야에 대한 세부 사항 검토를 국립국어원 주체로 외부 전문가의 자문을 얻어 작성함. 이에 포함되는 내용은 각 분야의 연구/개발할 세부내용, 결과물의 형태 및 평가지표, 지적소유권, 활용 방안, 개발 후 사후 운영/관리 방안, 과제 선정 절차 등이 있음.

○ 본 사업은 기존의 성과물(표준국어대사전, 세종계획 및 각 기관에서 수행한 결과물들) 등을 사업의 활용도 높은 성과를 얻기 위한 연계, 활용하여야 하며, 각 분야에서의 일관성과 체계성을 유지하는 데 노력해야 함.

○ 본 사업에서 유사성이 많은 부분(도구 개발, 협업 환경, 웹서비스 체계 구축 등)은 각 3개 분야에서 별도로 추진하지 않고 통합 추진하여 각 분야의 결과물 활용성을 높이고 공통된 부분에 대한 검증을 도모하며, 기존의 세종계획 결과물을 활용하여 일정한 공모전을 거쳐 선정하며 각종 도구 개발의 결과물로는 소스를 반드시 제출하도록 함.

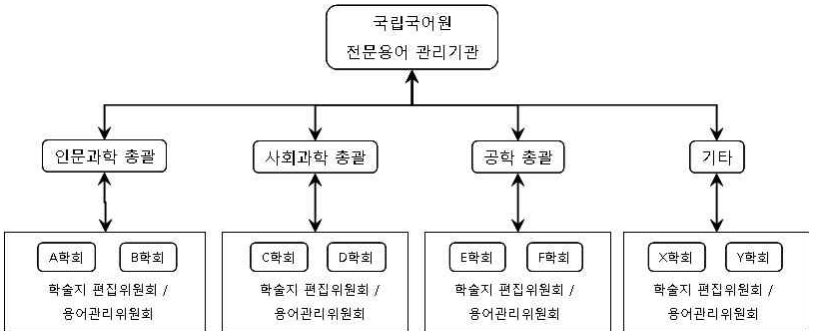
○ 사업을 통해 개발되는 결과물은 1차년도부터 매년 공개하고 홍보하여 사용자의 활용 활성화 및 참여를 유도하고, 사용자들로부터의 피드백을 통해 결과물의 신뢰도를 향상시키고 사업 기간 동안의 환경 변화에 적절하게 대응되도록 함.

○ 모든 결과물은 참여, 공유, 개방의 정신에 맞도록 웹 기반으로 제공 가능해야 하며, 특히 브라우저 종속적이지 않도록 표준 웹 가이드를 준수한 UI를 제공해야 함. 또한 관리 도구의 경우 오픈 소스화를 염두에 두고 표준화된 웹 기술로 구현되어야 할 것임.

○ 연구 성격이 강한 부분들은 각 분야에 일정 규모로 배분, 수행하되 이에 대한 가이드라인을 설정하여야 하고, 개발 성격의 결과물과는 평가지표 등을 달리하여 평가함.

○ 의미분석 자원 구축의 경우, 기존의 세종계획 결과물을 기반으로 하여 개발하며, 세종전자사전 및 표준국어대사전과도 연계되도록 함.

○ 본 사업 중 전문용어 분야는 대내외적으로 중요한 만큼 국립국어원에 국가전문용어센터 또는 이에 걸맞는 팀을 구성하는 것을 제안하며, 그 추진체계의 기본 개념은 아래 그림과 같으며, 각 학회에 표준적 가이드라인을 제시하고 적극 지원하여(예산 포함) 학회에서의 용어관리 위원회를 활성화 시킨다. (공동 협력 및 협조 체제 강화)



○ 국립국어원에서는 전문용어에 대하여 지속적으로 관리하여야 하고 이를 보급, 활성화하기 위한 각 부처와의 협력체계를 강화하여야 함. 기본적으로 국어기본법에 명시된 것을 적극적으로 적용해야 함.

1.2. 사업 평가 및 홍보

○ 각 사업에 대한 평가는 단계별로 하고, 평가는 국어심의회의 정보화 분과에서 함.

○ 연차별 최종평가는 별도의 전문가(한국정보과학회 언어공학연구회나 음성언어산업협회) 집단의 별도 평가위원회를 구성하여 매년 12월에 실시함.

○ 본 사업의 대내외적 협조 및 홍보를 위하여 매년 1회 이상 본 사업과 관련된 워크숍, 세미나(해외 전문가 참여) 등을 국립국어원 주관 하에 정기적으로 개최하여야 함.

2. 개발전략 및 방법

2.1. 지식사회를 선도하는 한국어

○ 연계 서비스 및 상용화를 위해서는 사업 초기부터 포털 사업자 및 언어처리 관련 기업체, 학교, 연구소를 주축으로 협의체를 구성하고 이를 중심으로 요구사항을 수렴할 필요가 있음. 협의체에서는 상용화를 위한 대상 영역, 언어 자원 규모, Raw Data 제공처 등 제반 사항을 결정하여 본 사업 수행 기관에 전달함. 과제 수행 기간 중에도 정기적으로 관련 정보를 교환하고 전문가 회의를 개최함으로써 과제 성공을 담보함.

○ 특정 플랫폼이나 H/W 환경에 종속되지 않도록 구현되어야 할 것임. 과제 전반에 걸쳐 H/W, S/W 등 주요 환경과 개발 언어, 개발 방법론을 표준화하여 추진해야 하며, 이러한 표준은 업계에서 사용하는 표준안을 준용함.

○ 연구-개발 과정에서 발생한 주요 연구결과물을 연구책임자 웹사이트를 통해 지속적으로 공개하여 연구의 진행상황 및 현재까지 결과물을 배포하고, 관련 연구자 및 추후 결과물 사용자들의 의견을 수렴하여 연구결과물의 활용가치를 극대화할 수 있도록 할 것임. 본 사업의 결과물을 이용하여 계속해서 응용 소프트웨어를 개발하거나 관련 분야의 연구에 결과물을 계속해서 활용하면서 오류를 발견하고 자체적으로 수정하여 업그레이드를 수행하며, 활용의 편의성 등을 고려하여 활용가치를 높여나가면서 유지보수를 가장 잘 수행할 수 있는 연구자가 사업을 주도적으로 수행해야 할 것임.

2.2. 미래를 준비하는 한국어

(1) 논항정보부착 말뭉치 구축

○ 언어학, 전산학자들로 구성된 논항정보태그 설계팀 구성

○ 태그부착 도구 설계 및 구현

○ 5만 어절 규모의 1차 시범 말뭉치 구축 및 이에 대한 문제점 파악

○ 시범 말뭉치 구축에서 드러나 문제점 보완 및 향후 말뭉치 구축 시 반영

○ 말뭉치 구축 초기단계부터, 말뭉치 검색 및 활용도구 개발 병행

○ 이에 기반한 대국민 베타서비스 준비

(2) 개체·시간·공간(ETS) 정보 부착 말뭉치 구축

○ 국어정보화 1단계 사업의 결과물 및 2단계 사업의 논항 정보 및 화행 정보 주석과의 호환성을 갖기 위해 공통 말뭉치를 선정하는 것이 바람직함.

○ 지식처리 영역에서 극복해야 할 문제점으로 대두되는 분산 텍스트 환경(다국어, 다문서, 대화)에서 추론을 하기 위해서, 연구 초기부터 국제적 협력 관계를 맺고 국제 표준안을 적극 수용함.

○ 국외에서도 ETS 정보 부착 연구는 지식 추론을 위한 소규모 시범 말뭉치 구축 단계임. 따라서 한국어 ETS 정보 부착 말뭉치의 경우도, 크기보다는 정확도와 활용 가능성 검토에 중점을 둠.

○ ETS 정보 부착 말뭉치를 구축하기 위해서는 1,2차년도에는 말뭉치 구축 자체보다 기초 연구에 주력함.

○ 산업계/학계 공청회를 통한 ETS 정보 분석 말뭉치 구축분야 선정하고, 언어학, 전산학자들로 구성된 ETS 정보 태그 설계팀 구성함.

○ ETS 정보 태그 부착 guideline과 도구 설계 및 구현

○ 10만 어절 규모의 1차 시범 말뭉치 구축 및 이에 대한 문제점을 보완하여 향후 말뭉치 구축 시 반영함.

○ 주석 말뭉치 구축이 완료되기 전, 말뭉치 검색, 평가 및 활용도구 개발 완료함.

○ 이에 기반한 대국민 베타 서비스를 준비함.

(3) 화행정보부착 말뭉치 구축

○ 언어학, 전산학자들로 구성된 화행정보태그 설계팀 구성

○ 태그부착 도구 설계 및 구현

○ 5만 어절 규모의 1차 시범 말뭉치 구축 및 이에 대한 문제점 파악

○ 시범 말뭉치 구축에서 드러나 문제점 보완 및 향후 말뭉치 구축시 반영

○ 말뭉치 구축 초기단계부터, 말뭉치 검색 및 활용도구 개발 병행

○ 이에 기반한 대국민 베타서비스 준비

(4) 논조정보부착 말뭉치 구축

○ 산업계/학계 공청회를 통한 논조분석 말뭉치 구축분야 선정

○ 언어학, 전산학자들로 구성된 논조정보태그 설계팀 구성

○ 포털서비스 전문가의 감수를 통한 각 분야별로 사용자들의 의견이 가장 많이 개진되는 영향력 있는 사이트 선정

○ 위 사이트에 게시된 문서로부터 논조정보 부착을 위한 문서 자동추출

○ 논조정보 부착을 위한 정련작업 (반자동 수행)

○ 태그부착 도구 설계 및 구현

○ 10만 어절 규모의 1차 시범 말뭉치 구축 및 이에 대한 문제점 파악

○ 시범 말뭉치에서 드러나 문제점 보완 및 향후 말뭉치 구축시 반영

○ 말뭉치 구축이 완료되기 전, 말뭉치 검색 및 활용도구 개발 완료

○ 이에 기반한 대국민 베타서비스 준비

(5) 다의어 의미부착 말뭉치 구축

○ 학계/산업계 공청회를 통한 세종1단계 의미체계 및 표준국어대사전 의미체계 통합방안 모색

○ 언어학, 국어학자들로 구성된 세종1단계 의미체계 및 표준국어대사전 의미체계 통합 태스크포스 구성

○ NLP관련 연구소/업체 전문가들이 요구하는 다의어 정보수준 분석

○ 태그부착 도구 설계 및 구현

○ 10만 어절 규모의 1차 시범 말뭉치 구축 및 이에 대한 문제점 파악

○ 시범 말뭉치에서 드러나 문제점 보완 및 향후 말뭉치 구축시 반영

○ 말뭉치 구축이 완료되기 전, 말뭉치 검색 및 활용도구 개발 완료

○ 이에 기반한 대국민 베타서비스 준비

(6) 호환성을 갖춘 한국어 어휘의미망 구축

○ 기개발된 중대형 규모 한국어 어휘의미망의 개발자를 중심으로 1차 개발팀을 구성하여 통합 어휘의미망인 I-KWordnet의 초기 개발기간을 단축함.

○ 기개발된 여러 어휘의미망의 상이한 강점을 상호 보완하여 통제된 기준을 사용하되, 어휘 간 풍요로운 의미관계를 제공할 수 있는 통합적 어휘의미망을 구현함.

○ 1차년도에 통합화된 I-KWordnet을 이용하여, 국어정보화 2단계 사업 초기부터 의미정보 부착 말뭉치에 분석 준거를 제공함으로써 말뭉치와 어휘의망 간의 연계성을 확보함.

○ 국어정보화 1단계 사업의 결과물인 세종전자사전과 국립국어원에서 주도하여 개발한 표준국어대사전을 I-KWordnet과 연계함으로써, 각 어휘의미가 제공할 수 있는 언어처리 정보의 단위를 다각화하고, 정보의 상세성을 심화함.

○ 현재 표준국어대사전의 뜻풀이 말에 나타난 문제점, 즉 어휘의 수와 수준이 통제되지 않고, 일관성이 결여되어 있다는 점을 어휘의미망의 상의-하의 관계를 이용하여 보완함.

○ 교차언어 검색, 기계번역 등 다국어 처리의 기초 자료를 제공할 수 있어야 함.

○ I-KWordnet의 구축/관리 도구 및 구조 정합성 검증 방법 개발

○ 이에 기반한 대국민 베타 서비스를 준비함.

2.3. 세계와 소통하는 한국어

(1) 전문용어 표준화 체계 구축

○ 전문용어의 기존 DB와 실제 검증을 하여야 하므로 이에 대응하는 관리팀, 개발팀, 검증팀 및 기관 간 협력체계가 가장 중요함.

○ 국외 표준에 대해서는 언어 독립적인 기술영역에 대하여 W3C, ETSI 등 참여하여 표준화 동향을 파악함.

○ 현재 전문용어 표현 양식에 국한하여 참여하고 있는 국외 표준화 활동을 확대하여 언어처리응용 기반기술인 언어분석 API 표준화와 언어처리응용엔진의 평가 체계 구축의 표준화에도 국내 표준화 포럼을 통해 적극 참여함.

(2) 언어자원관리 표준 및 생활 전문용어 구축

○ 언어자원 국제표준은 전문가 그룹을 형성화하기 위한 교육체계, 연구개발의 보편화와 보급을 전제로 함.

○ 언론사의 협조를 얻어 각 분야의 고빈도 생활 전문용어를 우선 선정하여 이에 대한 개념을 정비하고 한국어 IT HUB를 통해 서비스한다. 또한, 언론사에서 정의하는 생활 전문용어도 공유, 개방 되도록 함.

(3) 언어처리 기술 평가를 위한 표준화된 평가 세트 구축

○ 실제 정보검색 및 문서분류에 대한 평가 세트의 구축 작업은 방대한 문서의 양으로 인하여 매우 많은 인력 및 시간을 필요로 함. 특히 모든 문서를 수작업으로 처리하는 것은 거의 불가능함. 이러한 이유로 인하여 수동 색인은 전체 말뭉치의 일부에 대하여 수행하도록 할 수밖에 없음. 수동 태깅의 대상이 되는 부분 문서집합을 찾는 기술을 중심으로 연구하고 이를 활용하는 접근 기법이 필요함.

○ 정보검색 평가는 NTIS의 TREC, NTCIR의 현재 평가 토픽 및 방법을 따르되, 국내외적 평가를 위한 그룹을 초빙하여 개발하고 입력을 받는 절차가 필요함.

3. 결과물 서비스/홍보/사업화 방안

3.1. 지식사회를 선도하는 한국어

○ 개발된 콘텐츠와 결과물은 주로 인터넷을 통해서 활용하도록 함. 이미

자생적으로 콘텐츠가 증가하고 이의 수정 및 배포가 참여와 공유의 개념에 따라서 구축되게 되므로, 이러한 형태를 준용함. 개방형 플랫폼을 접근성이 용이한 포털 사이트와의 제휴를 통해 가장 효과적으로 홍보 및 사업화가 가능.

○ 포털 사이트와의 제휴는 1) 서비스 제공의 안정성을 확보하는 측면, 2) 실제 산업체에 적용됨으로써 인프라를 튜닝해 나가는 측면, 3) 지속적으로 실용어 Raw data를 확보하는 측면에서 의의를 가짐. 특히 포털 사이트와 적극적인 제휴 또는 공동 사업화를 통해서 국어IT플랫폼을 포털 사이트의 서비스 운용 인프라를 활용할 경우 서비스 운영의 안정성을 담보 받을 수 있게 될 것임.

○ 공동 사업을 추진하는 포털 사이트의 광고 영역을 통해 본 사업을 알림으로써 단기간에 많은 사용자를 확보하여 신속히 상용화에 돌입할 수 있게 될 것임. 공동 사업에 참여하는 포털 사이트에는 타사 대비 우선으로 언어 자원을 제공받는 특혜를 제공함으로써 사업 참여 동기를 부여.

○ 학술적 관점에서는 본 사업의 수행내역, 결과물 등을 연구책임자 홈페이지를 통해 지속적으로 연구결과물의 사용자들이 필요한 자료를 활용할 수 있도록 하여 정보가 빨리 확산될 수 있도록 함. 개인 연구자 및 비사업용으로 연구자에게는 연구결과물을 무료로 제공하고 산업체 등 사업화 관련 기관에는 연회비를 받아서 사업 종료후 연구결과물을 지속적으로 유지하는 비용으로 활용하는 방안을 모색하는 것도 좋은 방안임.

3.2. 미래를 준비하는 한국어

○ 말뭉치 검색 및 활용도구를 통해 학술적인 사용을 원하는 사용자들에게 무상 서비스 제공하며, 산업계는 유상 배포함.

○ 본 말뭉치를 대상으로 한 추론 시스템, Q&A 시스템 등의 contest를 개최함으로써 관련 연구자와 산업계에 선순환적인 경쟁을 유도함.

○ 말뭉치 구축이 완료되기 전, 음성/언어정보산업협의회 등에 샘플말뭉치를 배포하여 관련 업체들에 적극적인 홍보를 실시함.

○ 분야별 논조정보관련 기본어휘사전을 구축하게 되면 부가적인 언어자원 구축 및 사업화도 가능할 것으로 기대

○ 기계번역, 정보검색 등의 개발을 위해 의미모호성 모듈이 필요한 업체에는 유상으로 배포

○ I-KWordnet의 확장, 응용, 활용 시스템 등을 대상으로 contest를 개최함으로써 관련 연구자와 산업계에 선순환적인 경쟁을 유도함.

○ 어휘의미망은 산업계에서 가장 필요로 하는 언어자원인 만큼 산업계/학계 공청회를 통해 적극적인 의견 수렴하고 홍보함.

3.3. 세계와 소통하는 한국어

○ 전문용어의 경우 ISO의 Conceptual DB, 기술표준원의 용어 검색, 통계청의 검색창, 교육과학기술부의 교과서 편찬, 출판사 등의 협력이 필요하다. 국립국어원이 정부 기관 차원에서 통합된 체계 구축을 하도록 함.

o 전문용어의 사용자는 다양한 목적과 계층이 있기 때문에, One-source multi-service라는 차원에서 각각의 해당 기관들이 표준화된 용어를 각기 서비스하여 각각의 사용자들이 쓰도록 하는 정책이 우선되어야 함.

o 교과서, 웹, 언론 등의 매체에 대한 전문용어 인증제도를 실시 추천함.

o 정보검색 평가체계는 새로운 미래 인터넷을 대비하기 위한 것이므로 종래의 평가 대상보다는 블로그, 위키, 위키피디아, 음성대화 등과 같은 전략적 목표치를 미래에 두도록 함.

o 언어자원 국제화에 의하여, 한국어 연구를 국제적으로 집중되도록 매년 국제회의, 국제학술홍보, 국제특허검색, 기계번역 서미트와 같은 기회를 활용 홍보함.

3.4. 사업화 방안

(1) 기술이전 및 산업화를 위한 법・제도 정비

○ 구축 개발된 결과물(각종 언어자원 및 활용/지원 프로그램)에 대한 지적소유권 및 실시권에 대해 명확한 규정을 마련하고, 이의 언어처리 산업

체로의 기술이전이 가능하도록 관련 법·제도를 정비하여야 함.

○ 기술이전을 전담하는 전문가 팀을 구성하여 기술이전 및 산업화에 필요한 제반 업무를 담당함.

(2) 산·학·연·관 협력 체계 구축

○ 국어정보화 사업의 효율적인 대국민 서비스과 결과물의 산업화를 위해서는 산업계, 학계, 연구소 등의 역할을 명확히 분담함.

○ 특히 **한국어 IT HUB**의 구축 유지 보수를 위해 필요한 시스템 개발에는 사업 초기 단계부터 산업계가 참여할 수 있도록 함.

○ 전문용어는 전문 출판사와의 협업과 분업 관계의 체계를 구축하여야 함. 국립국어원에서는 국가 사전 관리 및 생산 방식을 국어정보화의 기술력을 발전시키는 방향으로 발전시키고, 전문용어를 민간 전문 출판사와 대학/연구소에 조건부로 공개하여 이를 개선하여 사업화하되 결과물을 국립국어원으로 환원(feedback)하여 전문용어가 확장/표준화되는 선순환 구조를 확립함.

(3) 국가 언어지식 지원 통합기기 개발

○ 언어지식 지원 통합기기는 전자북(e-book), 정보검색 서비스 등을 통합한 언어 정보 단말기로서, 모바일 상황에서도 경제, 정보통신 등 국민의 관심도가 높은 분야의 학술 및 생활 전문용어에 대한 개념을 제공하는 서비스가 가능함.

○ 아마존이 지난 2007년 11월에 'Kindle'이라는 전자북 단말기를 출시하면서 전자북 사용자 저변을 확대할 수 있는 다양한 방안을 검토하고 있음. 또한, 이를 계기로 전자북 시장에 새로운 활기를 불어 넣고 있음.

○ 국내에서도 모바일 환경에서의 정보검색이 점차 확산되는 추세에 맞추어 국내 연구소뿐만 아니라 기업체에서도 이러한 단말기를 개발하여 출시되고 있음.

○ 이러한 언어지식 지원 통합기기의 성공적인 개발을 위해서 출판사는 콘텐츠를 제공하고 포털은 서비스를 제공한다면, 국어정보화 사업 결과물의 대국민 서비스 및 사업화 수준을 한 단계 끌어올릴 수 있을 것임.

빈 페이지 임

VII. 소요 예산

1. 지식사회를 선도하는 한국어

(단위: 백만원)

세부 과제	1	2	3	4	5	합계
한국어 정보처리 인프라 구축을 위한 실용 언어 자원 구축	300	300	300	300	300	1,500
언어 산업과 연계한 언어 인프라 구축	250	250	250	250	300	1,300
국어 자원의 IT 활용을 위한 공유 체계 구축	300	360	400	450	400	1,910
(합계)	850	910	950	1,000	1,000	4,710

2. 미래를 준비하는 한국어

(단위: 백만원)

세부 과제	1	2	3	4	5	합계
다양한 의미분석 자원 구축	500	500	500	500	500	2,500
한국어 어휘의미망 구축	300	300	300	300	300	1,500
의미분석 자원 간의 연계/연동 시스템 구축	300	300	300	300	300	1,500
(합계)	1,100	1,100	1,100	1,100	1,100	5,500

3. 세계와 소통하는 한국어

(단위: 백만원)

세부 과제	1	2	3	4	5	합계
전문용어 정비 및 표준화 사업	300	300	300	300	300	1,500
언어자원 표준화 및 생활전문용어 구축	500	500	500	500	500	2,500
언어처리 기술 평가를 위한 표준화된 평가 세트 구축	300	300	300			900
(합계)	800	800	800	800	800	4,900

VII. 참고 문헌

- 웹사이트-

우리말 배움터: http://urimal.cs.pusan.ac.kr
코리안 클릭: http://www.koreanclick.com/
20Newsgroups: http://people.csail.mit.edu/jrennie/20Newsgroups/.
BalkaNet: http://www.ceid.uptras.gr/Balkanet/.
BNC(British National Corpus): http://www.natcorp.ox.ac.uk/
CLARIN project (유럽 언어자원 및 기술 공유화 사업), http://www.clarin.eu/
CoreNet: http://korterm.kaist.ac.kr/
EuroWordNet: http://www.illc.uva.nl/EuroWordNet/.
FlareNet project (유럽 언어자원네트워크 구축 사업),
http://www.ilc.cnr.it/flarenet/
GermaNet: http://www.sfs.uni-tuebingen.de/lsd/.
GNU: www.gnu.org
Google AdSense: http://www.google.com/adsense.
Google Blog: http://googleblog.blogspot.com/
GWC(세계워드넷 연합): http://www.globalwordnet.org/gwa/wordnet_table.htm.
ISO/TC37, http://www.iso.org/tc/
ISO/TC37/SC4, http://www.tc37sc4.org/
KorLex: http://corpus.fr.pusan.ac.kr/korlex/start.htm.
LISA(Localization Industry Standards Association) : www.lisa.org
Memodata: http://www.memodata.com.
NTCIR: http://research.nii.ac.jp/ntcir/
NTICR 일본 아시아언어 정보검색 평가사업, http://research.nii.ac.jp/ntcir/
PWN: http://wordnet.princeton.edu.
Sourceforge: http://web.sourceforge.com/
Technorati: www.technorati.com
TREC 미국 정보검색 평가사업, http://trec.nist.gov/
TREC(Text REtrieval Conference) : http://trec.nist.gov/
U-Win: http://nlplab.ulsan.ac.kr/

-한국어-

국립국어연구원. 2002. 『현대 국어 사용 빈도 조사: 한국어 학습용 어휘 선정을 위한 기초 조사』, 국립국어원.
국립국어연구원. 2005. 『현대 국어 사용 빈도 조사2』, 국립국어원.
국립국어원. 2001. 『표준국어대사전 1.0』, 두산동아.
김양진. 2006. "국어 중사전의 전문어 표제어 선정에 대하여",『한국사전학』7, 191-215.
김종만 외. 2007. "인터넷 정보량 급증의 영향과 대응", 삼성경제연구소.
김지영, 장동현, 맹성현, 이석훈, 서정현, 김현. 2000. "한국어 테스트 컬렉션 HANTEC의 확장 및 보안", 제12회 한글 및 한국어 정보처리 학술대회, 210-215.
맹성현, 이석훈, 이준호, 이응봉, 송사광. 1999. "정보검색시스템 평가를 위한 균형 테스트 컬렉션 구축", 한국정보관리학회지, 제6권, 제2호.
문유진. 1996. 『의미론적 어휘 개념에 기반한 한국어 명사 워드넷의 설계와 초록』, 서울대학교 컴퓨터공학과 박사학위 청구논문.
서정연. 1993. 『대화체 기계번역에 대한 연구』, 한국통신 연구보고서.
애틀러스 리서치 앤 컨설팅. 2007. "포스트 Google의 세계적 조류와 차세대 검색엔진의 진화 방향". 애틀러스 리서치
애틀러스 리서치 앤 컨설팅. 2008. "집단지성의 사업화에 도전하는 구글과 위키피디아, 'Knol' vs. 'Wikia Search'". 애틀러스 리서치
옥철영. 2007. "어휘의미망과 국어사전의 체계적 구성", 『한국어 어휘의미망 구축과 사전편찬 학술회의 자료집』, 국립국어원, 35-53.
윤애선 외. 2009. "한국어 어휘의미망 KorLex 1.5의 구축", 『한국정보과학회 논문집』, 2009년 1월 출판 예정.
윤애선. 2007. "국내·외 어휘의미망의 구축과 활용", 『새국어생활』 17(3), 5-25.
이성헌, "사전편찬에 있어서의 어휘의미망의 역할과 기능", 『한국어 어휘의미망 구축과 사전편찬 학술회의 자료집』, 국립국어원, 77-90.
이창기, 이근배. 2000. "의미애매성 해소를 이용한 WordNet자동 매핑", 『제12회 한글 및 한국어정보처리 학술대회 발표논문집』, 262-268.
임재현 외. 2008. "Web 2.0시대의 인터넷 사업 성공 요건", LG경제연구소.
정국. 1993-5.『한국어 특질에 대한 연구』, 한국통신 연구보고서.
정영임, 조선호, 윤애선, 권혁철. 2008. "구문 관계와 운율 특성을 이용한 한국어 운율구 경계 예측", 『인지과학』, vol. 19, no. 1, pp. 89 -105.
최경봉, 도원영. 2005. "한국어 동사 의미망 구축을 위한 상위 온톨로지 구성에 관한 연구", 『한국어학』 28, 217-244.
최기선 외. 2005. 『다국어 어휘의미망(CoreNet)』 3 vols., 한국과학기술원 전문용어언어공학연구센터, KAIST Press.
최재웅. 1996. "대화분석에 있어서의 몇가지 문제: 호텔 예약 전화대화를 중심으로", 1996년 인지과학회 학술대회 자료집.
최호섭 외. 2006. "대규모 우리말 어휘지능망 구축 방법", 『한글』 273, 125-141.
홍재성 외. 2007. 『21세기 세종계획 전자사전 개발 연구보고서 (11-1370252-000063-10)』, 문화관광부.

-영문-

Ahrenberg, L.,N. Dahlback and A. Jonsson. 1995. "Codings Schemas for Studies of Natural Language Dialogue". *Working Notes from the AAAI Spring Symposium*.

Alexandersson, J.,B. Buschbeck-Wolf,T. Fujinami,M. Kipp,S. Koch,E. Maier,N. Reithinger,B. Schmitz and M. Siegel. 1998. Dialogue Acts in Verbmobil-2 (Second Edition). In *Verbmobil Report 226*. Saarbrucken: DFKI.

Allen, J. and M. Core. 1997. Damsl: Dialogue Act Markup in Several Layers (Draft 2.1). In *Technical Report (Multiparty Discourse Group: Discourse Resource Initiative)*.

Allen, J.F. 2005. "Towards a General Theory of Action and Time", *The Language of Time*, I. Mani, J. Pustejovsky, R. Gaizauskas (eds.), Oxford University Press, 251-276.

Allwood, J.,J. Nivre and E. Ahlsen. 1994. Semantics and Spoken Language Manual for Coding Interaction Management. In *Report from the HSFR project Semantik och talsprak*.

Alsina, A. 1996. *The Role of Argument Structure in Grammar*. Center for the Study of Language and Information, Stanford, CA.: CSLI Publications, Stanford University.

Austin, J.L. 1976. *How to Do Things with Words (Second Edition)*. London: Oxford University Press.

Bach, E. 1981. "On Time, Tense, and Aspect: An Essay in English Metaphysics", P. Cole (eds.), 63-81.

Bach, K. and R.M. Harnish. 1979. *Linguistic Communication and Speech Acts*. Cambridge, MA: MIT Press.

Bain, J. 2006. "Spacetime Structualism", *The Ontology of Spacetime*, D. Dieks (ed.), Elsevier, 37-65.

Barker, C. and D. Dowty. 1993. "Non-Verbal Thematic Proto-Roles". *NELS* 23:49-62.

Bittar, A. 2008. "Annotation des informations temporelles dans des textes en fraçais", in *Proceeding of RECITAL(Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues)*, Avignon, 9-13 join 2008. http://www.lia.univ-avignon.fr/index.php?id=644.

Bunt, H. 1995. "Dynamic Interpretation and Dialogue Theory". In *The Structure of Multimodal Dialogue*, D. Bouwhuis & F. Neel M.Taylor (eds.), 139-166.

Amsterdam: John Benjamins.

Bunt, H. 2000. "Dialogue Pragmatics and Context Specification". In *Abduction, Belief and Context in Dialogue. Studies in Computational Pragmatics*, H. Bunt & W. Black (eds.), 81-150. Amsterdam: John Benjamins.

Bunt, H. 2006. "Dimensions in Dialogue Act Annotation". *Proceedings LREC 2006 Workshop*.

Bunt, H. 2007. "Multifunctionality and Multidimensional Dialogue Act Annotation". In *Communication - Action - Meaning*, E. Ahlsen et al. (eds.), 237-259: Gothenburg University.

Bunt, H. 2007. "The Semantics of Semantic Annotation". *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation (PACLIC-21)*. 13-28.

Bunt, H. 2008. Language Resource Management ? Semantic Annotation Framework ? Part 2: Dialogue Acts. Ms., *ISO/WD 24617-2 (Draft)*.

Bunt, H. and A. Schiffrin. 2006. "Methodological Aspects of Semantic Annotation". *Proceedings LREC 2006*.

Busemann, S., Decleck, T., Diagne, A. K., Dini, L., Klein, J. and Schmeier, S. 1997. "Natural Language Dialogue Service for Appointment Scheduling Agents", *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 25-32.

Carletta, J. 1996. "Assessing Agreement on Classification Tasks: The Kappa Statistic". *Computational Linguistics* 22:249-254.

Choi, K.-S. 1995. "Proceedings of Natural Language Processing Pacific Rim Symposium '95."

Clark, H.H. 1996. *Using Language*. Cambridge, UK: Cambridge University Press.

Core and Allen. 1997. "Coding Dialogs with the Damsl Annotation Schema". *AAAI Fal Symposium on Communicative Action in Humans and Machines*.

Dau, F., Mugnier, M.L., & Steumme, G. (eds.). 2005. *Conceptual Structures: Common Semantics for Sharing Knowledge*, Springer.

Di Eugenio, B.,P.W. Jordan and L.Pylkkanen. 1998. The Coconut Project: Dialogue Annotation Manual. In *ISP TEchnical Report 98-1*. University of Pittsburgh.

Dong, Z, & Dong, Q. 2006. *HowNet and the Computation of Meaning*, World Scientific.

Earman, J. 2006. "The Implications of Genral Covariance for the Ontology and Ideology of Spacetime", *The Ontology of Spacetime*, D. Dieks (ed.),

Elsevier, 3-23.
Edmondson, W. 1981. *Spoken Discourse*. London: Longman.
Evens, M.W.(ed.) 1988. *Relational Models of the Lexicon*, Cambridge University Press, Cambridge.
Fellbaum, Ch. (ed.) 1998. *WordNet: An Electronic Lexical Database*, The MIT Press, Cambridge.
Ferro, L., Gerber, L., Mani, I., Sundheim, B. & Wilson G. 2005. *TIDES 2005 Standard for the Annotation of Temporal Expressions.* April 2005 the September version updates all references to the ISO 8601 standard to incorporate the latest edition (8601:2004, Third Edition). http://fofoca.mitre.org/annotation_guidelines/timex2_annotation_guidelines.html
Galton, A. 2005. "A Critical Examination of Allen's Theory of Action and Time", *The Language of Time*, I. Mani, J. Pustejovsky, R. Gaizauskas (eds.), Oxford University Press, 277-300.
Geertzen, J. and H. Bunt. 2006. "Measuring Annotator Agreement in a Complex, Hierarchical Dialogue Act Schema". *Proc. SIGDIAL 2006*.
Gildea, D. and D. Jurafsky. 2002. "Automatic Labeling of Semantic Roles". *Computational Linguistics*.
Godfrey, J.,E. Holliman and J. and McDaniel. 1992. "Switchboard: Telephone Speech Corpus for Research and Development". *Proc. ICASSP*:517-520.
Gross, M. 2002. " Les déterminants numéraux, un exemple : les dates horaires", *Langage* 145, 21-38.
Han, Chung-Hye, Na-Rare Han, Eon-Suk Ko, Martha Palmer. 2002. "Development and evaluation of a Korean Treebank and its application to NLP," *Proceedings of the 3rd International Conference on Language Resources and Evaluation 2002*.
Harman D. 1993. "Overview of the 1st text retrieval conference", Proc. of 16th ACM SIGIR, 36-48.
Harman, D., Voorhees, E. & Buckland, L. "Overview of the TRECs - 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16", Proceedings of TREC, NIST Special Publication. also available at http://trec.nist.gov/pubs.html.
Hitzeman, J. 2007. "Text type and the Position of a Temporal Adverbial Within the Sentence", *Lecture Notes in Artificial Intelligence 4795* (Revised Papers of the International Seminar on Annotating, Extracting and Reasoning About Time and Events, held in Dagstuhl Castle, Germany, April 2005), Springer, 29-40.
Hobbs, J. & Pan, F. 2004. "An Ontology of Time for the Semantic Web", *ACM Transaction on Asian Language Information Processing* 3(1), pp.66-85.
Hovy, E. 2005. "Methodologies for the Reliable Construction of Ontological Knowledge", *LNAI* vol.359, 91-106.
Hwang, S.H., Y.I. Jung, A.S. Yoon, H.C. Kwon. 2006. *Building Korean Classifier Ontology Based on Korean WordNet*. TSD-LANI .
Hwang, S.H., Yoon, A.S. & Kwon, H.C. 2008. "Semantic representation of Korean numeral classifier and its ontology building for HLT applications", *Language Resources and Evaluation 42-2,* 151-172.
Ikehara, S. & al. 1997. *The Semantic System, vol. 1 of Goi-Taikei, A Japanese Lexcion*, Iwanami Shoten.
ISO-TimeML Working Group. 2008. ISO CD 24617-1, *Language Resource Management: Semantic Annotation Fremework Part1: Time and Events*, ISO/TC37/SC4 N412 rev02 (published on 15th Aug, 2008).
Jefferson, G. 1984. "Notes on a Systematic Deployment of the Acknowledgement Tokens 'Yeah' and 'Mm Hm'". *Papers in Linguistics* 17:197-216.
Jekat, S.,A. Klein,E. Maier,I. Maleck,M. Mast and J.J. Quantz. 1995. Dialogue Acts in Verbmobil. In *Verbmobil Report 65*. Saarbrucken: DFKI.
Jelinek, E. and A. Carnie. 2003. "Argument Hierarchies and the Mapping Principle". In *Formal Approaches to Function in Grammar: In Honor of Eloise Jelinek*, H. B. Harley A. Carnie, and M. Willie (eds.), 265-296. Amsterdam: John Benjamins.
Joachims, T. 1998. "Text categorization with support vector machines: learning with many relevant features", In Proc. of ECML '98, 137-142.
Jung, Y., A.S. Yoon, H.C. Kwon. 2007. "Grapheme-to-Phoneme Conversion of Arabic Numeral Expressions for Embedded TTS Systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1. 296-309.
Jung, Y.I., A.S. Yoon, H.C. Kwon. 2006. "Disambiguation Based on WordNet for Transliteration of Arabic Numerals for Korean TTS", *Proceedings of the 7th International Conference on Intelligent Text Processing and Computational Linguistics CICLings*, 366-377
Jurafsky, D.,E. Schriberg and a.D. Biasca. 1997. Switchboard Swbd-Damsl Shall-Discourse-Function Annotation Coders Manual. In *Technical Report 97-02*. Boulder, CO: University of Colorado, Institute of Cognitive Science.
Kang, M., S. Jung, K. Park & H.C. Kwon. 2007. "Part-of-Speech Tagging Using

Word Probability Based on Category Patterns", *Lecture Notes in Computer Science*, Vol. 4394. 119 ~ 130.

Katz, G. & Arosio, F. 2005. "The Annotation of Temporal Information in Natural Language Sentences", *The Language of Time*, I. Mani, J. Pustejovsky, R. Gaizauskas (eds.), Oxford University Press, 513-522.

Katz, G. 2007. "Towards a Denotational Semantics for TimeML", *Lecture Notes in Artificial Intelligence 4795* (Revised Papers of the International Seminar on Annotating, Extracting and Reasoning About Time and Events, held in Dagstuhl Castle, Germany, April 2005), Springer, 88-99.

Kim, J.A.,Y.H. Cho,J.W. Lee and G.C. Kim. 1995. "A Response Generation in Dialogue System Based on Dialogue Flow Diagrams". *Proceedings of NLPRS '95.*

Ko, Y. & Seo, J. 2004. "Learning with unlabeled data for text categorization using bootstrapping and feature projection techniques", In Proc. of the ACL-04.

Koppel, M. 2007. "Measuring Differentiability: Unmasking Pseudonymous Authors", *Journal of Machine Learning Research 8,* 1261-1276

Kwon, H.C., M. Kang, S. Choi. 2004. "Stochastic Korean Word Spacing with Smoothing Using Korean Spelling Checker," *Computer Processing of Oriental Languages*, vol. 17. 239-252.

Lee, E.R., Yoon, A.S. & Kwon, H.C. 2007. "Exploiting Morpho-syntactic Features for Verb Sense Distinction in KorLex", *ICCS 2007, Lecture Notes in Computer Science 4488*, 1170-1177.

Lee, H. 2003. *Prominence Mismatch and Markedness Reduction in Word Order.*vol. 21: Natural Language and Linguistic Theory.

Leech, G. 2004. "Adding Linguistic Annotation". In *Developing Linguistic Corpora:A Guide to Good Practice*, Martin Wynne (eds.). Oxford: Oxbow Books.

Levin, B. and M.R. Hovav. 2005. *Argument Realization*. Cambridge, MA: MIT Press.

Levinson, S.C. 1983. *Pragmatics*. Cambridge, UK: Cambridge UniversityPress.

Lewis, D. 1997. "Reuters-21578 text categorization test collection README file", Manuscript.

Lewis, D., Yang, Y., Rose, T. & Li, F. 2004. "RCV1: A New Benchmark Collection for Text Categorization Research", J of Machine Learning Research, Vol. 5, 361-397.

Mani, I., Pustejovsky, J. & Gaizauskas, R. (eds.) 2005. *The Language of Time*, Oxford University Press.

Meteer, M. and A. Taylor. 1995. Dysfluency Annotation Stylebook for the Switchboard Corpus. Ms., *Distributed by LDC.*

Moulton, L. 2008. "Enterprise Search Markets and Applications Capitalizing on Emerging Demand", *The Gilbane Group Research report.*

Nirenburg, S. & Raskin, V. 2004. *Ontological Semantics*, The MIT Press.

Noriko K., Kazuko K., Toshihiko N., Koji E., Hiroyuki K., Soichiro H., & Jun A. 1999. "The NTCIR Workshop: the First Evaluation Workshop on Japanese Text Retrieval and Cross-Lingual Information Retrieval", Proc. of IRAL '99.

Pala, K., & Sedláček R. 2005. "Enriching WordNet with Derivational Subnets", *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, 305-311.

Paul, S. 2007. "Recommending related papers based on digital library access records", *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, 417-418*

Petukhova, V.V. 2005. Multidimensional Interaction of Multimodal Dialogue Acts in Meetings.

Popescu-Belis, A. 2005. "Dialogue Acts: One or More Dimensions?" *ISSCO Working Paper* 62.

Pustejovsky, J. & Verhagen, M. 2008. "Implementing TimeML-Based Applications for Temporal Reasoning", *Proceedings of 2008 PNU International Conference on Language and Knowledge Processing*, 85-92.

Pustejovsky, J., Littman, J. & Saurí, R. 2007. "Arguments in TimeML: Events and Entities", *Lecture Notes in Artificial Intelligence 4795* (Revised Papers of the International Seminar on Annotating, Extracting and Reasoning About Time and Events, held in Dagstuhl Castle, Germany, April 2005), Springer, 107-126.

Sacks, H.,E.A. Schegloff and a.G. Jefferson. 1974. "A Simplest Systematics for the Organization of Turn-Taking for Conversation". *Language* 50-4:696-735.

Saurí, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A. & Pustejovsky, J. 2006. *TimeML Annotation Guidelines Version 1.2.1.*

Schalley, A., & Zaefferer, D. (eds.). 2007. *Ontolinguistics: How Ontological Status Shapes the Linguistic Coding of Concepts*, Mouton de Gruyter.

Schegloff, E. 1968. "Sequencing in Conversational Openings". *American Anthropologist* 70:1075-1095.

Schegloff, E. and H. Sacks. 1973. "Opening up Closings". *Semiotica* VIIII:289-327.

Schiffrin, D. 1987. *Discourse Markers*. Cambridge, UK: Cambridge University Press.

Schilder, F. & Habel, C. 2005. "From Temporal Expression to Temporal Information: Semantic Tagging of News Messages", *The Language of Time*, I. Mani, J. Pustejovsky, R. Gaizauskas (eds.), Oxford University Press, 533-545.

Schilder, F., Katz, G. & Pustejovsky, J. 2007. "Annotating, Extracting and Reasoning About Time and Events", *Lecture Notes in Artificial Intelligence 4795* (Revised Papers of the International Seminar on Annotating, Extracting and Reasoning About Time and Events, held in Dagstuhl Castle, Germany, April 2005), 1-6.

Searle, J.R. 1969. *Speech Acts*. Cambridge, UK: Cambridge University Press.

Searle, J.R. 1983. *Intentionality*. Cambridge, UK: Cambridge University Press.

Sebastiani, F. 2002. "Machine learning in automated text categorization", ACM Computing Surveys, 34(1), 1-47.

Seo, J.Y.,J.W. Lee,J.H. Kim,J.M. Cho,C.H. Kim and a.G.C. Kim. 1994. "Dialogue Machine Translation Using a Dialogue Model". *The First China-Korea Joint Symposium on Machine Translation*.

Soria, C. and V. Pirrelli. 2003. "A Multi-Level Annotation Meta-Schema for Dialogue Acts". In *Computational Linguistics in Pisa: Linguistica Computazionale Special Issue*, N. Calzolari & L. Cignoni A. Zampolli (eds.), 865-900. Pisa-Roma: IEPI.

Sowa, J. 1999. *Knowledge Representation: Logical, Philisophical, and Computational Foundations*, Brooks and Cole.

Stolcke, A.,K. Ries,N. Coccaro,E. Shriberg,R. Bates,D. Jurafsky,P. Taylo,R. Martin,C.V. Ess-Dykema and M. Meteerr. 2000. "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech". *Computational Linguistics* 26-3:339-373.

Traum, D. 2000. "20 Questions on Dialogue Act Taxonomies". *Journal of Semantics* 17-1:7-30.

Traum, D. and E. Hinkelman. 1992. "Conversation Acts in Task-Oriented Spoken Dialogue". *Computational Intelligence* 3-8:575-599.

Traum, D. and S. Larsson. 2003. "The Information StateApproach to Dialogue Act Management". In *Current and New Directions in Discourse and Dialogue*, J. van Kuppevelt & R. Smith (eds.). Dordrecht: Kluwer.

Verhagen, M., Mani, I., Saurí, R., Knippen, R., Littman, J. & Pusterjovsky, J. 2005. "Automating Temporal Annotation with TARSQI", In *Proceedings of the ACL 2005*.

Verkuyl, H. 1972. *On the compositional nature of the aspects*, Reidel, Dordrecht.

Verkuyl, H. 1993. *A Theory of Aspectuality. The Interaction between Temporal and Atemporal structure*, Cambridge Univ. Press.

Vossen, P. 1998. *EuroWordNet: A Multilinugal Database with Lexical Semantic Network*, The Kluwer Academic Publishers.

Weber, E.G. 1993. *Varieties of Questions in English Conversation*. Amsterdam: John Benjamins.

Yablonsky, S., & Sukhonogov, A. 2006. "Semi-Automated English-Russian WordNet Construction", *Proc. of the 3rd Int'l WordNet Conference*, 345-347.

Yoon, A. S. et al. 2003. "An Automatic Transcription System for Arabic Numerals in Korean", *Proceedings of 2003 International Conference on Natural Language Processing and Knowledge Engineering*. 221~226.

Yoon, A. S. et al. 2004. "Automatic Transcription of Three Ambiguous Symbols Used with Arabic Numerals: Period, Colon and Slash ", *Language and Information*, Vol. 8, Jun. 2004. 117-136.

연구자 주소록

연구책임자: 서영훈(충북대학교 전기전자컴퓨터공학과 교수)
e-mail : yhseo@chungbuk.ac.kr
공동연구원: 강승식(국민대학교 컴퓨터공학부 교수)
e-mail : sskang@kookmin.ac.kr
김경선(다이퀘스트 연구 소장)
e-mail : kksun@diquest.com
윤애선(부산대학교 불어불문학과 교수)
e-mail : asyoon@pusan.ac.kr
최재웅(고려대학교 언어학과 교수)
e-mail : jchoe@korea.ac.kr
최호섭(한국과학기술정보연구원 정보기술개발단 연구원)
e-mail : hschoe@kisti.re.kr
자문위원 : 강현규(건국대학교 컴퓨터응용과학부)
e-mail : yhseo@chungbuk.ac.kr
심철민(파란닷컴 본부장)
e-mail : ailove1@paran.com
박동인(한국과학기술정보연구원 e-science본부)
e-mail : dipark@kisti.re.kr
옥철영(울산대학교 컴퓨터정보통신공학부)
e-mail : okcy@ulsan.ac.kr
이성헌(서울대학교 불어불문학과)
e-mail : lsh0717@snu.ac.kr
최기선(한국과학기술원 전자전산학부)
e-mail : kschoi@cs.kaist.ac.kr
홍문표(성균관대학교 독어독문학과)
e-mail : skkhmp@skku.edu

국어정보화 2단계 사업 계획 수립

발행인 : 사단법인 한국정보과학회
발행처 : 국립국어원
서울시 강서구 방화3동 827
전화 02-2669-9775
인쇄일 : 2008년 12월 20일
발행일 : 2008년 12월 22일
인쇄처 : 성진인쇄
전화 052-277-4596

국립국어원
2008-01-21

국어정보화 2단계 사업 계획 수립

국립국어원

이 보고서는 국립국어원에서 시행한
"국어정보화 2단계 사업 계획 수립" 연구 용역 결과
보고서입니다.

THE NATIONAL INSTITUTE OF KOREAN LANGUAGE