

OSS 과제 수행: 개요

2010. 05. 19

KAIST 최기선 교수님 연구실

1. 연구목표 및 내용

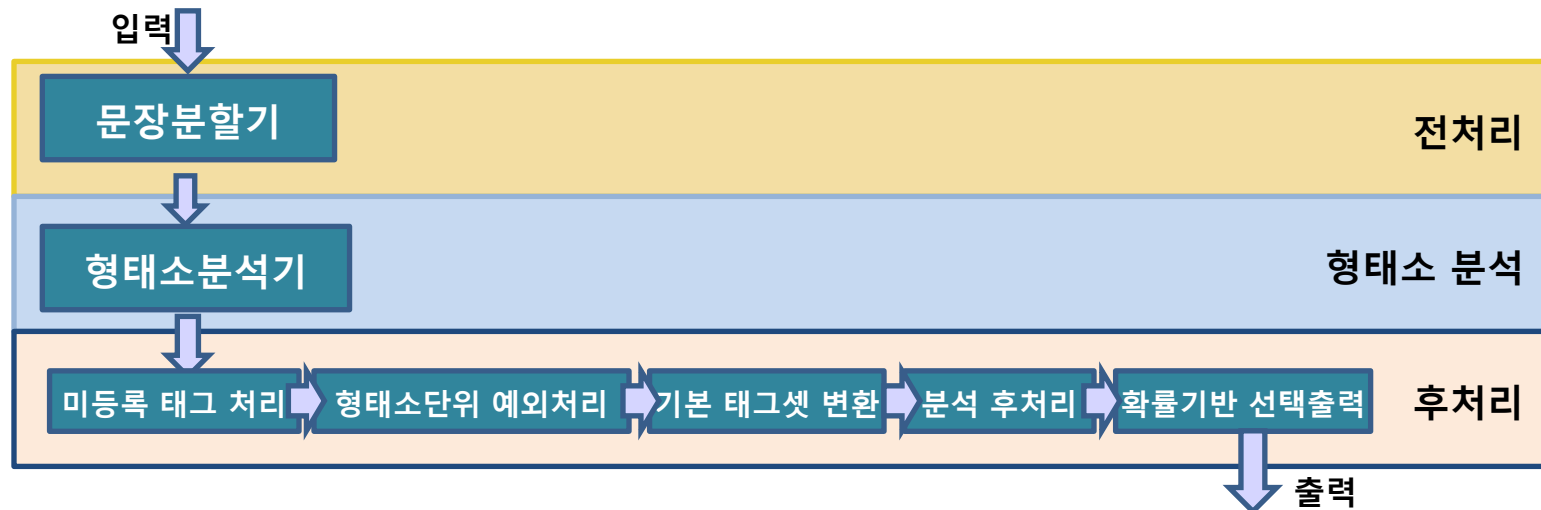
연구 목표

- 한국어 언어 자원 처리 및 접근 도구 개발
 - 한국어 형태소 분석기 Hannanum의 개량
 - 한국어 반자동 시맨틱 어노테이션 환경의 조성
 - 주요 한국어 자원 CoreNet에 대한 접근성을 높이는 툴킷 개발 및 커뮤니티 조성

목표 1: Hannanum의 개량 (1/2)

■ Hannanum 이란?

- 오픈소스 형태소 분석 라이브러리
- 한나누미(한국어 형태소 분석기) + 외래음새기미(영-한 음차 표기 모듈)로 구성
- 공개:
 - <http://kldp.net/projects/hannanum/>



목표 1: Hannanum의 개량 (2/2)

■ Hannanum의 문제점

- Linux Kernel에서 C로 개발됨: Platform-dependent
 - ✓ 현재 Windows 버전은 존재하지 않음
- 기반 문서의 부족: 초보자가 사용하기 힘들

■ 개량에 대한 올해 목표

- 현 Hannanum을 Java 로 전환
 - ✓ Platform independent, 타 프로그램에 이식 용이
- 기반 기술 문서의 작성, 설명서 작성
- 기타 기능 개량
 - ✓ 음운규칙, 품사관리, 사전관리 등을 모듈화하고 사용자 접근을 용이하게 함

목표 2: 한국어 반자동 시맨틱 어노테이션 환경 조성 (1/3)

■ 시맨틱 어노테이션:

- 문서로부터 사용자가 요구하는 정보를 추출해내는 기술
- 자동화를 위해서는 대규모 코퍼스가 필요하나, 현재 이 코퍼스를 구축하기 위한 지원 툴은 태부족

[3D0 게임기]는 **제작** 로닉
아츠]에서 출자하여 [3D0]라
는 회사를 만들어 **소유** 제작한
[32비트 게임기]이다. 게임
산업의 발달로 규모가 점점
커짐에 따라 거대 가전기업
들이 그 사업에 뛰어들었고
[파나소닉], [LG], [삼성], [산
요] 등의 거대 가전기업과
[EA]에서는 게임시장에서
새로운 [닌텐도] **프로젝트 실행**
있는 게임시장에 뛰어들고
자 [3D0프로젝트]를 실행했
다....

목표 2: 한국어 반자동 시맨틱 어노테이션 환경 조성 (2/3)

■ 관련 연구

➤ 국내

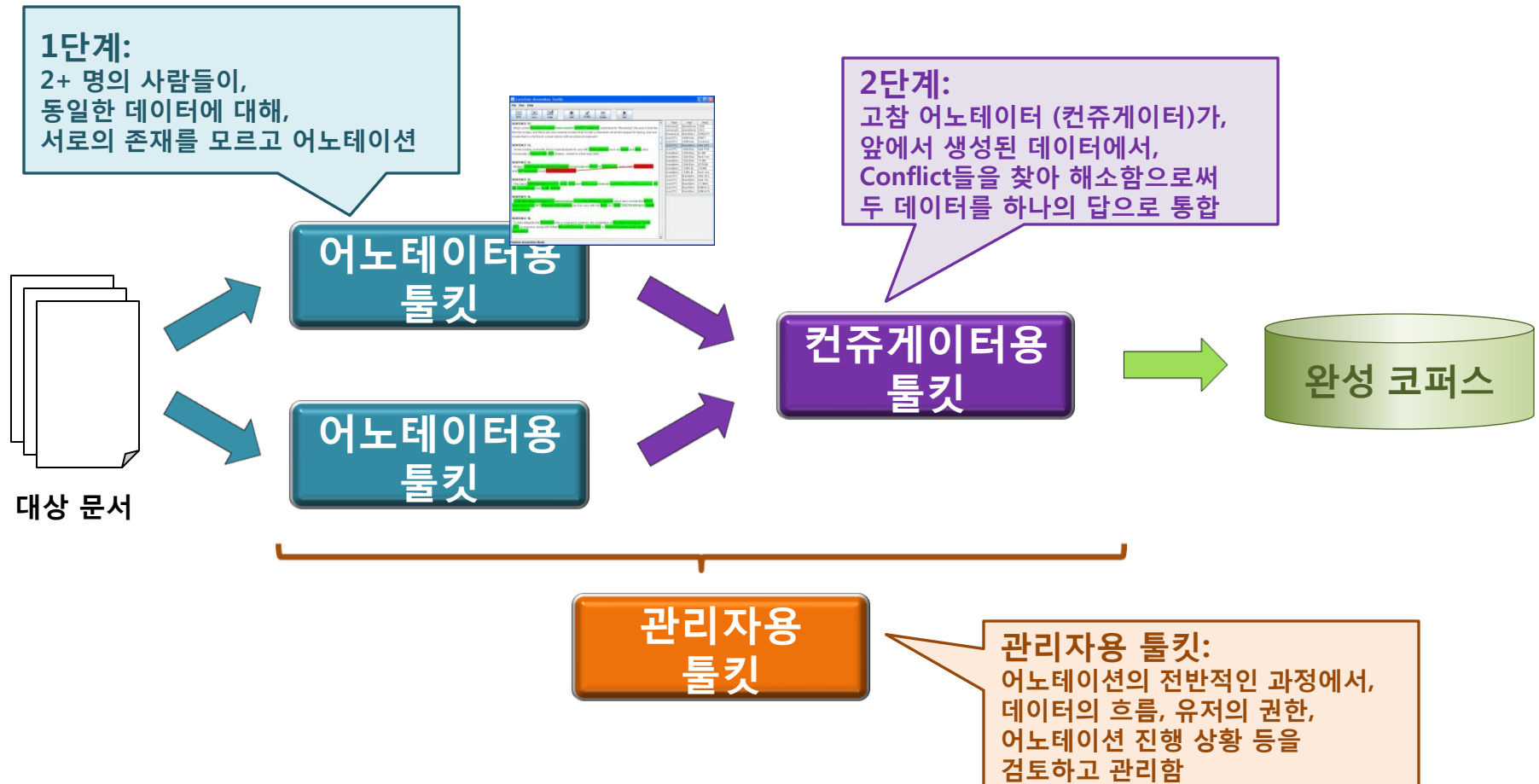
- 영남대의 OntoCS 등이 있으나, 공개된 툴킷은 없음.

➤ 국외

- Kim Platform(<http://www.ontotext.com/kim/docs.html>)
 - 엔티티 어노테이션만 지원
- XConc Suite(<http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=XConc+Suite>)
 - GENIA 프로젝트의 일환으로 개발된 eclipse plug-in
 - 엔티티, 관계 및 이벤트 어노테이션을 모두 지원하나, 컨쥬게이션 기능 부재
- LDC ACE Annotation Toolkit(<http://projects.ldc.upenn.edu/ace/tools/>)
 - 현재 관리 및 업데이트되고 있지 않음
 - 공개된 버전은 작동되지 않음
 - 데이터를 파일로 주고받기 때문에 전체적인 데이터 흐름 조율 및 관리가 힘들

목표 2: 한국어 반자동 시맨틱 어노테이션 환경 조성 (3/3)

■ 개요도



목표 3: CoreNet 접근성 높이기 (1/2)

■ CoreNet이란?

- 한국어를 중심으로 한 한중일다국어 개념망으로 일본 NTT 어휘 대계를 기반으로 구축된 2937개의 개념을 중심으로 명사, 동사, 형용사의 3만개 이상의 어휘가 6만여개의 의미로 연결되어진 언어망.
- 특징
 - 7000만 어절 코퍼스로부터 추출된 어휘
 - 한,중,일 다국어간 개념 체계 공유
 - 명사, 동사, 형용사로 이루어진 종합적 체계
 - 정의문

한국어	어휘번호	의미번호	개념번호	개념명	일본어	중국어
사과	1	0	11322531	과일 I:C:水果 IJ:果物	アップル, 林 子, れいし, レイシ, リン ゴ, リンゴ, 林檎	
사과	2	0	111131224	기타 관료 I:C:官吏(其他) IJ:役人 (その他)		
사과	3	0	11322531	과일 I:C:水果 IJ:果物	アップル, 林 子, れいし, レイシ, リン ゴ, リンゴ, 林檎	苹果
사과	5	0	122113272	신중 I:C:謹慎 IJ:慎み		
사과	6	1	1211112	학문분야/학과 I:C:学问領域/课程 IJ: 学問分野・学科		
사과	7	0	1211123	깨달음 I:C:理解 IJ:悟り		
사과	8	0	122126324	사죄 I:C:謝罪 IJ:謝罪	あやまり, 謝 り, 多謝, 平 謝り	歉意, 道歉

동사	개념번호	어휘 의미 번호 번호	N1				N2				N3			
			개념명	개념번호	논항	격조사	개념명	개념번호	논항	격조사	개념명	개념번호	논항	격조사
가다		0 0												
가다		0 0	인간	11111		이/가	결혼	122123441	장가	을/를				
가다		0 0	인간	11111		이/가	여자	11111212	시집	을/를				
가다		0 0	인간	11111		이/가					학교	112114	학교	에/에게/로
가다		0 0	직업적인 일	1221282	일	에	손/팔	11311231	손	이				
가다		0 0	인간	11111		이/가					집회	12123341	모임	에/에게/로
가다		0 0	인간	11111		이/가					회	1112215	동창회	에/에게/로
가다		0 0	인간	11111		이/가	지역(인간 활동)	11222	장소	을/를				
가다		0 0	상품	1132212	물건	에	손/팔	11311231	손	이				
가다		0 0	식료	113225	식료	에	손/팔	11311231	손	이				
가다	122225911	0 1	인간	11111		이/가	관람	122118118	구경	을/를				
가다	122225911	0 1	인간	11111		이/가	이사	12212236	이민	을/를				
가다	122225911	0 1	인간	11111		이/가	이사	12212236	이사	을/를				
가다	122225911	0 1	인간	11111		이/가	견임	12212323	견근	을/를				
가다	122225911	0 1	인간	11111		이/가	재학	122123312	유학	을/를				
가다	122225911	0 1	인간	11111		이/가	여행	12212411	여행	을/를				
가다	122225911	0 1	인간	11111		이/가	도망	122225631	도망	을/를				
가다	122225911	0 1	인간	11111		이/가	도망	122225631	망명	을/를				
가다	122225911	0 1	인간	11111		이/가	나가기	12222611	출장	을/를				
가다	122225911	0 1	할것(본체 (명))	11322911	차	이/가	사회	11124	세상	에				
가다	122225911	0 1	인간	11111		이/가								
가다	122225911	0 1	인간	11111		이/가					오른쪽	123882	오른쪽	에/에게/로
가다	122225911	0 1	인간	11111		이/가					영화관	1121153	영화관	에/에게/로

목표 3: CoreNet 접근성 높이기 (2/2)

■ 목표: CoreNet 검색기 구현

➤ 코어넷의 복잡한 정보를 쉽게 검색할 수 있도록 인터페이스 구성

- 개념 체계의 트리 구조 탐색 지원
- 개념 및 명사의 키워드 기반 검색
- 동사 및 형용사 격률 정보의 검색
- 개념 중심의 명사, 형용사, 동사 동시 탐색 지원
- 우리말 큰사전의 정의문 제공

2. 추진 일정

추진 일정

과제내용	추진 일정											
	1	2	3	4	5	6	7	8	9	10	11	12
Hannanum 이식: Java Platform												
Hannanum: 기능 개선 (음운규칙, 품사관리, 사전관리 모듈)												
Hannanum: 통합 테스트												
어노테이션 툴킷: 어노테이터용 툴킷 및 통합용 툴킷 개발 및 개량												
어노테이션 툴킷: 관리자용 툴킷 개발												
어노테이션 툴킷: 자동화 프로토타입 구축												
어노테이션 툴킷: 자동화 시스템 안정화												
CoreNet: GUI 툴킷 개발												
주요 Milestone 완성점에서의 수행결과	- 기능 개선된 한나눔 공개 - 어노테이션 툴킷 1차 공개 - 개발보고서 및 기술문서						- 음차표기 모듈 통합된 한나눔 - 반자동 어노테이션 툴킷 공개 - 개발보고서 및 기술문서					

감사합니다!