한국어 언어 자원 처리 및 접근 도구

- 중간 실적 보고

KAIST

2010. 09. 15

목 차

1. 목표대비 개발진행상황

2. 커뮤니티별 활동내용 및 주요성과
   a. 한국어 형태소 분석기 '한나눔'
   b. 시맨틱 어노테이션 도구 'COAT'

1. 목표대비 개발진행상황

수행계획서 상 목표

과제내용	추 진 일 정											
	1	2	3	4	5	6	7	8	9	10	11	12
Hannanum: 기능 개선												
Hannanum: 기능 확장												
Hannanum: 통합 테스트												
어노테이션 툴킷: 어노테이터용 툴킷 및 통합용 툴킷 개량												
어노테이션 툴킷: 관리자용 툴킷 개발												
어노테이션 툴킷: 자동화 프로토타입 구축												
어노테이션 툴킷: 자동화 시스템 안정화												
~~CoreNet: GUI 툴킷 개발~~												

수행계획서 작성시, 9월 15일 시점에 도출될 것으로 계획되었던 성과물

- 한국어 형태소 분석기 HanNanum
  - 기능 개선 및 확장된 형태소 분석기: 모듈화 및 각 모듈의 수정 지원
  - 기술 문서: 매뉴얼
- 어노테이션 툴킷
  - 어노테이터용 툴킷, 통합용 툴킷 및 관리자용 툴킷
  - 기술 문서: 매뉴얼

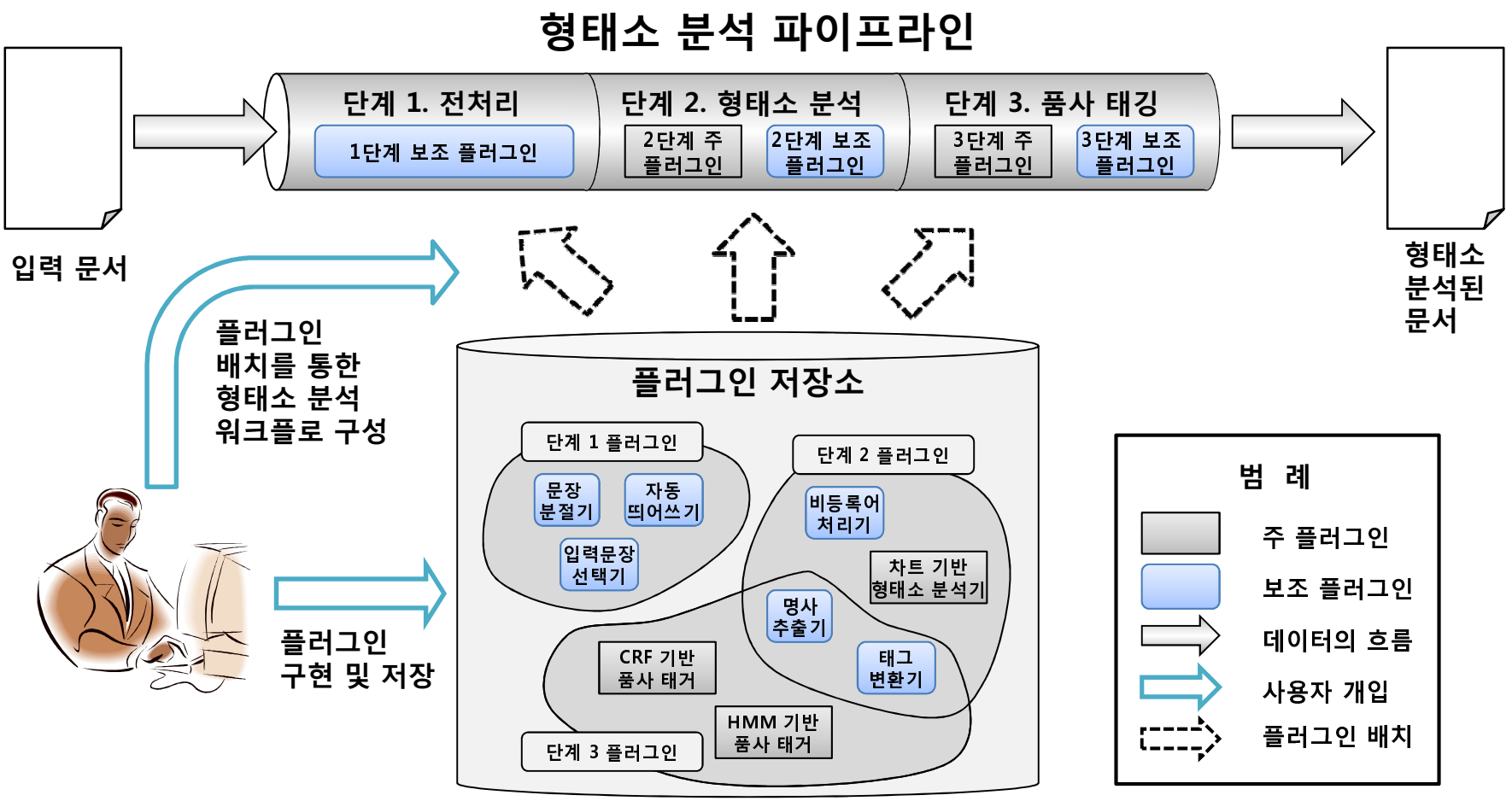
수행계획서 대비 진행 상황

- 수행계획서 작성시, 9월 15일 시점에 도출될 것으로 계획되었던 성과물
  - 한국어 형태소 분석기 HanNanum
    - 기능 개선 및 확장된 형태소 분석기: 모듈화 및 각 모듈의 수정 지원
      - → HanNanum의 손쉬운 사용 및 다른 도구와의 통합을 위하여, Java 버전으로 재구현
      - → 각 모듈의 입/출력을 정확히 정의하고, 새로운 사용자로 하여금 자신의 상황에 알맞도록 모듈을 재구현하여 자신만의 형태소 분석 Workflow를 구축할 수 있도록 지원
    - 기술 문서: 매뉴얼
      - → 작성중
  - 어노테이션 툴킷
    - 어노테이터용 툴킷, 통합용 툴킷 및 관리자용 툴킷
      - → 어노테이터용 툴킷, 통합용 툴킷 및 관리자용 툴킷의 구현 완료
      - → 어노테이터용 툴킷, 통합용 툴킷 및 관리자용 툴킷을 하나로 모은 COAT Package V0.5를 공개
    - 기술 문서: 매뉴얼
      - → 작성 및 공개 완료 (V 0.5)

2. 커뮤니티별 활동내용 및 주요성과

a. 한국어 형태소 분석기 '한나눔'

한나눔 개념도



각 모듈의 입출력을 정확히 정의하고, 사용자로 하여금 재구현할 수 있는 JAVA 인터페이스를 제공함으로써, 형태소 분석기의 확장성 및 범용성의 극대화를 꾀함

개발 진행 상황

- **기반 시스템 구축**
  - 다중 쓰레드 기반 파이프라인 상에서 플러그인들을 동작시키는 작업 완료
  - 각 단계의 플러그인 컴포넌트 간 명확한 입 · 출력을 정의하는 작업 진행 중
  - 편리한 성능 평가를 위한 모듈 개발 구상 중

- **컴포넌트 플러그인 개발**
  - 현재까지 개발된 플러그인 :
    - Lattice 형태의 Chart 기반 형태소 분석기, HMM 기반 품사 태거, 문장 분리기, 비형식적 문서 입력 필터, 미등록어 처리기
  - 개발 계획중인 플러그인 :
    - CRF 기반 품사 태거, 대용량 코퍼스 기반 형태소 분석기, 자동 띄어쓰기 처리기, 명사 추출기, 태그 변환기

- **문서화**
  - 사용자가 개발된 시스템을 쉽게 사용할 수 있도록 안내하는 매뉴얼 작성 중
  - 시스템 구성, 파이프라인 설정 방법, 플러그인 개발 방법, 개발된 플러그인에 대한 설명, 형태소 사전, 태그셋 등의 내용 포함

커뮤니티 활동

- 활동 커뮤니티
  - http://kldp.net/projects/hannanum

- 업데이트 현황 (2010년 9월 9일 기준)
  - 현재 릴리즈 버전 jhannanum 0.7.4 (자바 버전)
  - 8월 이후 총 릴리즈 업데이트 횟수 5회
  - 누적 다운로드 횟수: 122회

  - 현재 SVN 리비전 66
  - 8월 이후 SVN 커밋 횟수 6회

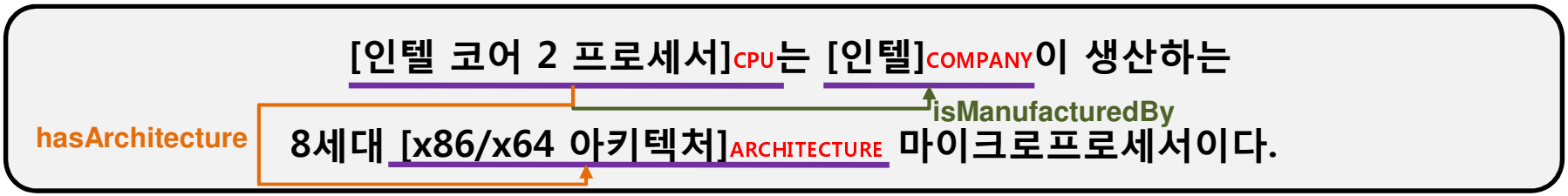
2. 커뮤니티별 활동내용 및 주요성과

b. 시맨틱 어노테이션 도구 'COAT'

개요

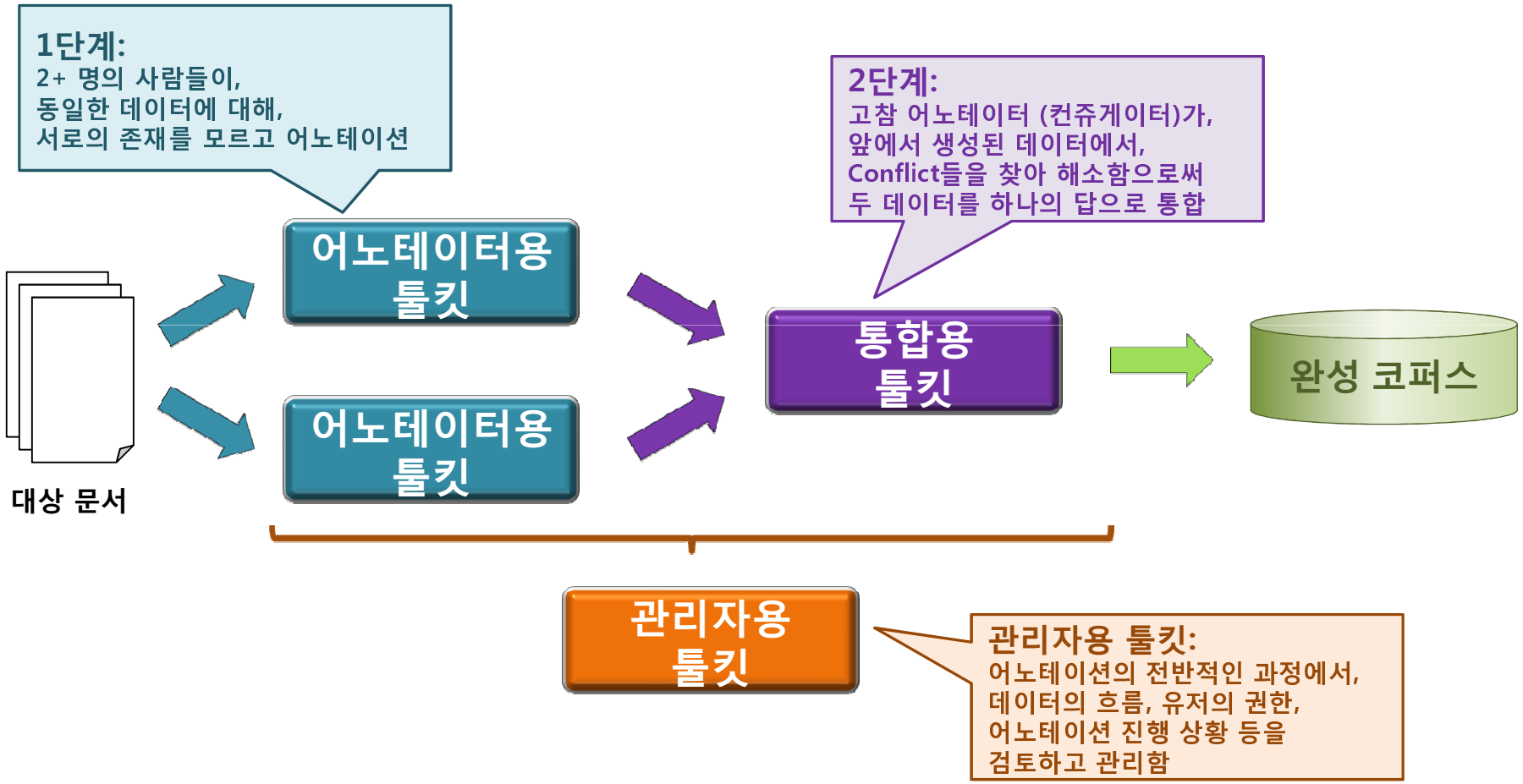
- 어노테이션이란?
  - 자연언어처리 분야에서, 어노테이션이란 원시 코퍼스에 대한 메타데이터(meta-data)를 의미한다(출처: Wikipedia).
  - 일반적으로, 기계 학습 방식으로 자연언어처리 모듈을 구축하기 위해,
    - ➢ 먼저 어느 정도 규모의 원시 코퍼스에 대하여 목적으로 하는 자연언어처리 모듈의 이상적인 처리 결과를 어노테이션하고,
    - ➢ 이후 어노테이션된 데이터를 훈련 및 검증 데이터로 사용함으로써 목표로 하는 자연언어처리 모듈을 구축한다.

- 시맨틱 어노테이션이란?
  - 원시 코퍼스에서 각 단어의 의미, 또는 각 단어간에 존재하는 의미적인 관계를 나타내는 메타데이터를 의미한다.
  - 시맨틱 어노테이션의 예시:



  - 시맨틱 웹의 활성화를 위해서는, 기존 문서에 자동으로 시맨틱 어노테이션을 수행하는 기능이 꼭 필요하다.

COAT 개념도



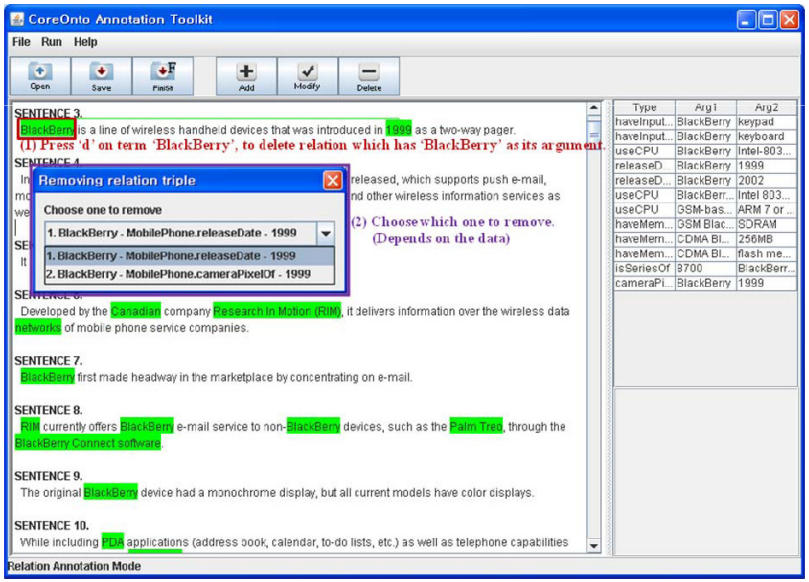
COAT의 특징

1. 데이터의 신뢰성을 높이기 위하여, 두 명 이상의 사람이 동일 문서에 대하여 작업한 후, 작업된 문서를 통합하는 방식 사용

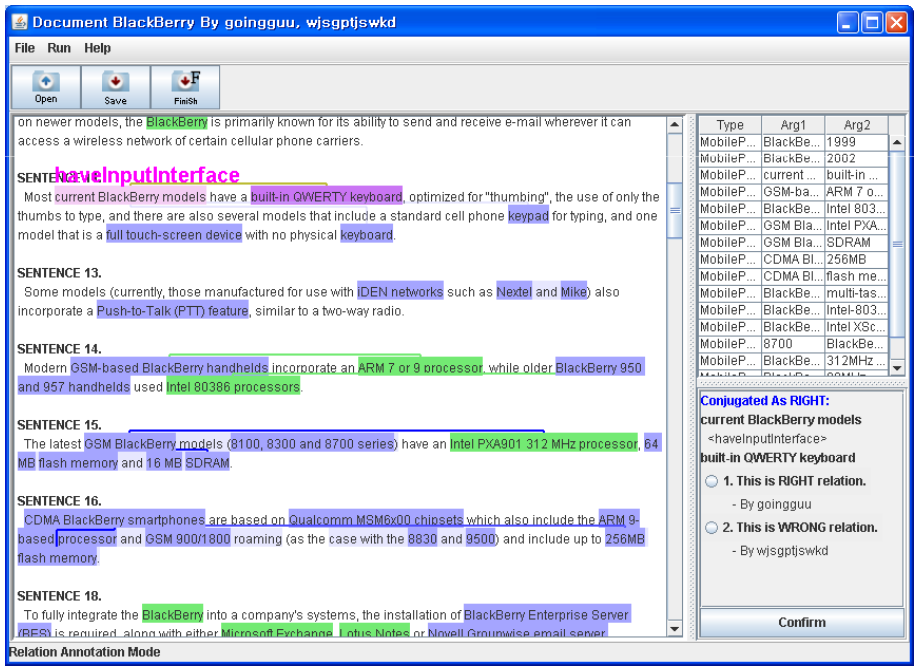
2. 작업 전체 과정의 진행 및 데이터의 사용을 쉽고 빠르게 하기 위하여, 작업의 진행도를 한눈에 확인할 수 있는 관리자용 도구의 개발

COAT 인터페이스

- 원시 코퍼스 상에 의미 정보가 그림으로 표현되도록 지원
- 실제 적용시 어노테이션 비용이 1/10으로 절감 되었음



<COAT Annotator>



<COAT Conjugator>

커뮤니티 현황 및 향후 계획

- 홈페이지:
  - http://sourceforge.net/projects/coatsemantic/
  - 2010년 5월 3일 개설
  - 현재까지 누적 다운로드 횟수: 134회

- 향후 계획
  - 사용자가 구축한 데이터를 토대로 시맨틱 어노테이션을 자동화하는 모듈의 프로토타입 개발