CognIT a.s

KunDoc –
Knowledge-based Document analysis

ScandSum meeting

Åre

March 19th, 2004

Till C. Lech

till@cognit.no

KunDoc

This Presentation

- KunDoc – an Introduction
- Overview – State of the Art in Co-reference chaining
- Overview – Ontologies and Semantic Web technology
- KunDoc – Working Hypotheses/Methodology

KunDoc – Fact Sheet

- Research Project, funded by the Norwegian Research Council (KUNSTI)
- Duration: 3 years, started in Oct. ’03
- Partners: University of Bergen, CognIT as
- Goal: Develop a method for Knowledge-based Co-reference chaining
- Dissemination: Publications, PhD-thesis, Work Shops, Demonstrator

KunDoc – Background and Motivation

- Recognition of co-reference chains is essential for tasks that rely on a thorough semantic document analysis:
  - Summarisation
  - Information Extraction
  - Information Retrieval
- Knowledge-based methods often fail due to the lack of available knowledge

KunDoc – Background and Motivation

- Knowledge Management and Semantic Web Initiatives have developed tools for handling large scale knowledge life cycles.
- Can language technology and KM-technology be combined in order to improve co-reference chaining?

Co-reference Chaining – Anaphora Resolution

- Overview: Mitkov (1999) State of the art Report
- Traditional vs alternative approaches:
  - Traditional: Discount unlikely candidates, keep the one with highest salience value
  - Alternative: Most likely candidate is computed on basis of traditional models

Co-reference chaining – Anaphora Resolution

- Few knowlegde-based approaches: Wilks' Preference Samantics; earlier: Frames/Scripts
- Recent trends: Statistical methods (BREDT-project), Hybrid methods (Stuckardt: RoSANA-ML)

The Semantic Web – Background

- Information on WWW structured according to layout only (HTML)
- High quality IR/IE dependant on semantic annotation of Web Content
- Goal: Make Web Content processible by machines (agents)
- Storage and Interchange of Knowledge shall be enabled by means of formal Ontologies

The Semantic Web – Achievements

- Standards for Representation of content developed:
  - XTM, RDF, RDF(S), DAML+OIL, OWL
  - Formalisms for Representation and Interchange of ontologiese
- Several tools developed for acquisition and maintenance of knowledge bases
  - Text mining: Text-to-Onto (AIFB)
  - Extraction: OnTo-Extract (CognIT)
  - Editing: Protege2000 (Stanford), Onto-Edit (FZI)

The Semantic Web – Results so far

- Theoretical Framework for Knowledge Representation and Queries
- Tools to handle the Framework
- Ontologies developed in many domains

KunDoc and Ontologies – some central Questions

- What kind of domain knowledge is needed in order to support co-reference chaining?
- How can this knowledge be acquired and used?
- How can formal ontologies provide the framework needed for Co-reference chaining?

KunDoc – Approach

- Transformation of syntactic/statistic knowledge into explicit domain knowledge

Knowledge Extraction

- Hypothesis: Verbs found in documents define the relations between concepts an a given domain
- Occurence of verbs define the possible relationships
- Statistics determine the predominant relations

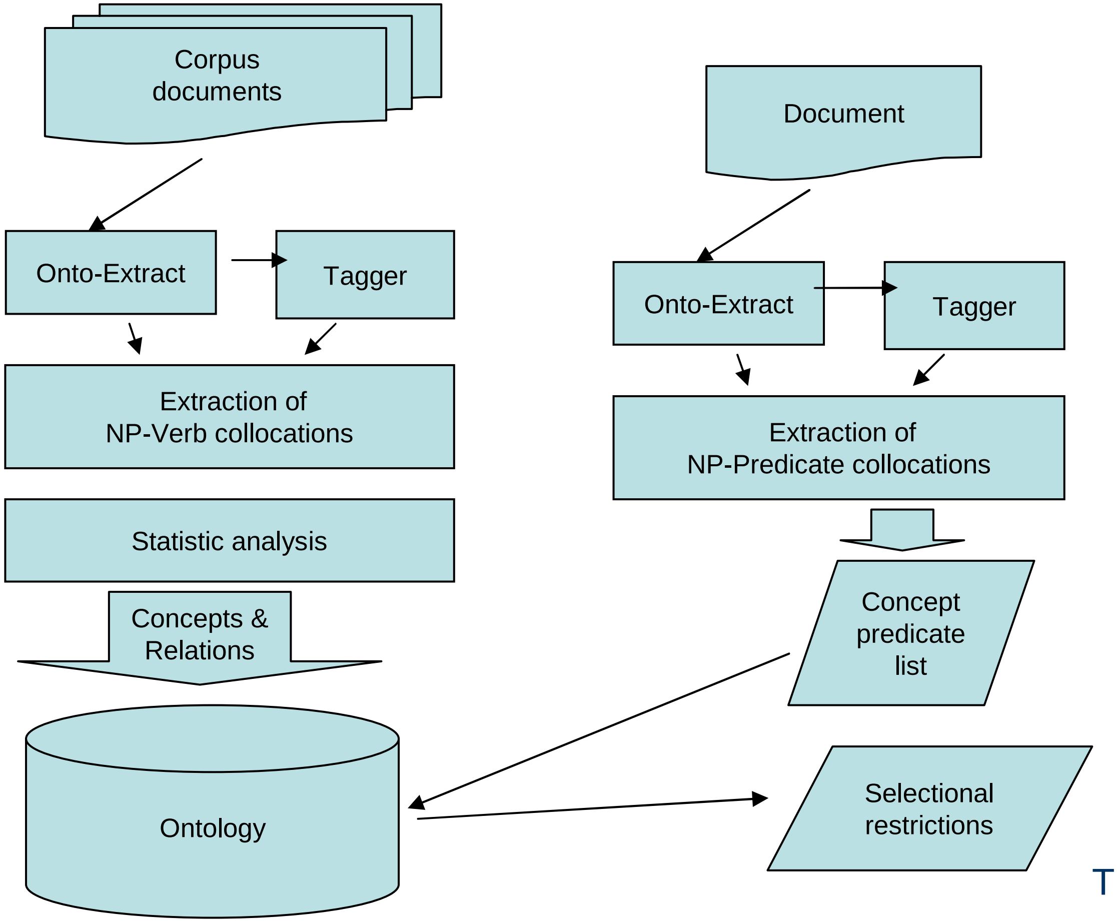
Knowledge Extraction

- Two parallell approaches
- UiB: Using precise parsing results from NorGram/XLE parser to extract NPs and predicates
- CognIT: Fuzzy approach, using Concept-verb collocations

Methods and Tools available

- CORPORUM Onto-Extract, for semantic annotation
- Oslo-Bergen Tagger
- Norgram-Grammar, XLE Parser
- Statistic models – unsupervised learning for verb clustering, latent semantic analysis
- Knowledge representation: RDF(S), Protégé2000

Architecture – Draft



KunDoc – Impact

- Improved Co-reference chaining will improve the analysis of text and thus improve related applications:
  - Summarisation
  - Information Extraction
  - Information Retrieval
  - Text mining
- Methods developed in KunDoc will also improve the automatic extraction of knowledge bases