Knowledge Representation from Information Extraction

Tan Xu, [1,3] Douglas W. Oard, [1,3] Tamer Elsayed, [2,3] Asad Sayeed [2,3]
[1]College of Information Studies, [2]Computer Science Department and [3]UMIACS CLIP Lab
University of Maryland, College Park, MD 20742, USA
{tanx, oard, telsayed, asayeed}@umd.edu

Categories and Subject Descriptors
I.2.7 **[Artificial Intelligence]:** Natural Language Processing

General Terms: Design, Experimentation, Standardization

Keywords: Information extraction, knowledge representation

1. INTRODUCTION
Digital libraries naturally occupy a middle ground between unstructured information (e.g., documents) and structured information (e.g., metadata). Information extraction techniques offer the potential to help bridge this divide by extracting structured content from unstructured sources in ways that support more complex reasoning than would otherwise be possible. In our research, we are exploring the potential to extend existing techniques for information extraction and within-document co-reference resolution with new techniques for cross-document co-reference resolution in ways that are designed to support ontological reasoning at collection scale.

2. IMPLEMENTATION

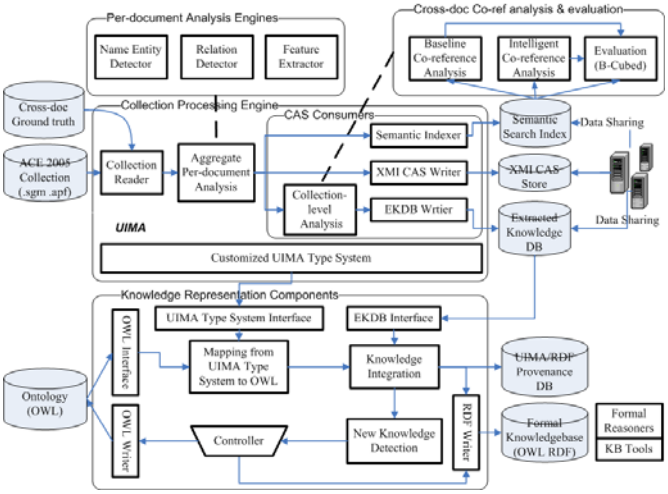

Figure 1 Prototype System Architecture

The architecture of our prototype system is shown in Figure 1. We are using ground truth annotation of entities and within-document co-reference for the English ACE 2005 training corpus as our experimental data; we ultimately plan to apply these techniques to the results of automatic extraction and within-document co-reference for the ACE 2008 evaluation collection [1]. Our information extraction components are embedded in the Unstructured Information Management Architecture (UIMA), an open-source middleware platform for integrating components that analyze unstructured information [2]. We produce a knowledge base from text in a pipeline that proceeds through the following stages:

JCDL'08, June 15-20, 2008, Pittsburgh, Pennsylvania, USA.
ACM 978-1-59593-998-2/08/06.

Per-document Preprocessing. The purpose of this stage is to assign annotations to regions of text, and to perform analysis on each document independently to extract more information that we need for further computing. The analysis includes feature extraction, named entity recognition, and relationship detection. The results are captured in a CAS (Common Analysis Structure, as defined by the UIMA type system). At the end of this process, we build an XMI CAS (XMI representation of the CAS) store and a semantic index that provides application-specific access to the CAS content.

Cross-document Co-reference Analysis. The annotations produced in the first stage are used as input for collection-level processing, the most important of which for our purpose is co-reference analysis—identification of identical entities that are mentioned in different documents. Our main approach to conflating entities across documents is to leverage evidence from the local, topical and social context of each mention [4]. We have also implemented a simple baseline where we conflate mention pairs that exhibit a small Levenshein edit distance between their respective heads. We use the B-cubed measure to compare these approaches [5].

Knowledge Integration. Since the UIMA type system permits no more than single-inheritance type/subtype hierarchies, to support substantive reasoning we must convert CAS results into a more expressive representation. We first identify the mapping of types in the UIMA type system to classes and properties in an OWL (Web Ontology Language) "target ontology," and then construct an RDF graph that instantiates the target ontology. This provides the opportunity to use existing reasoning engines already developed for OWL representations to perform more sophisticated deductive search. This process is similar to related work at IBM [3].

Ontology Expansion. Our initial target ontology, which consists of 165 classes and 63 properties, was generated using hand-annotated ACE 2005 documents. We found that although this ontology can cover most extracted entities, events, relations, values, time expressions, mentions and supporting concepts, but that we still lack targets for some of the contents of those ACE Program Format (APF) annotations. Thus, we plan to extend our current ontology to cover a broader range of the types that might be extracted.

3. REFRENCES
[1] Automatic Content Extraction (ACE) Evaluation. http://www.nist.gov/speech/tests/ace/

[2] Ferrucci, D. & Lally, A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering 2004.* 10 (3/4): 327–348.

[3] Ferrucci, D., Murdock, J.W. and Welty, C. Knowledge Integration in UIMA (a.k.a. SUKI). IBM Research Report RC24074. 2006.

[4] Elsayed, T., Oard, D.W. and Namata, G. Resolving Personal Names in Email Using Context Expansion. *ACL/HLT*, 2008.

[5] Bagga, A. & Baldwin, B. Algorithms for Scoring Coreference Chains. *Message Understanding Conference*, 1998.