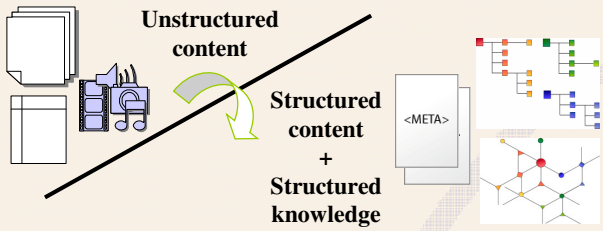
University of Sheffield NLP

Danica Damljanović, Florence Amardeilh, Kalina Bontcheva

KCAP'09, Redondo Beach, CA

CA Manager Framework: Creating Customised Workflows for Ontology Population and Semantic Annotation

Main Objective: bridging the gap



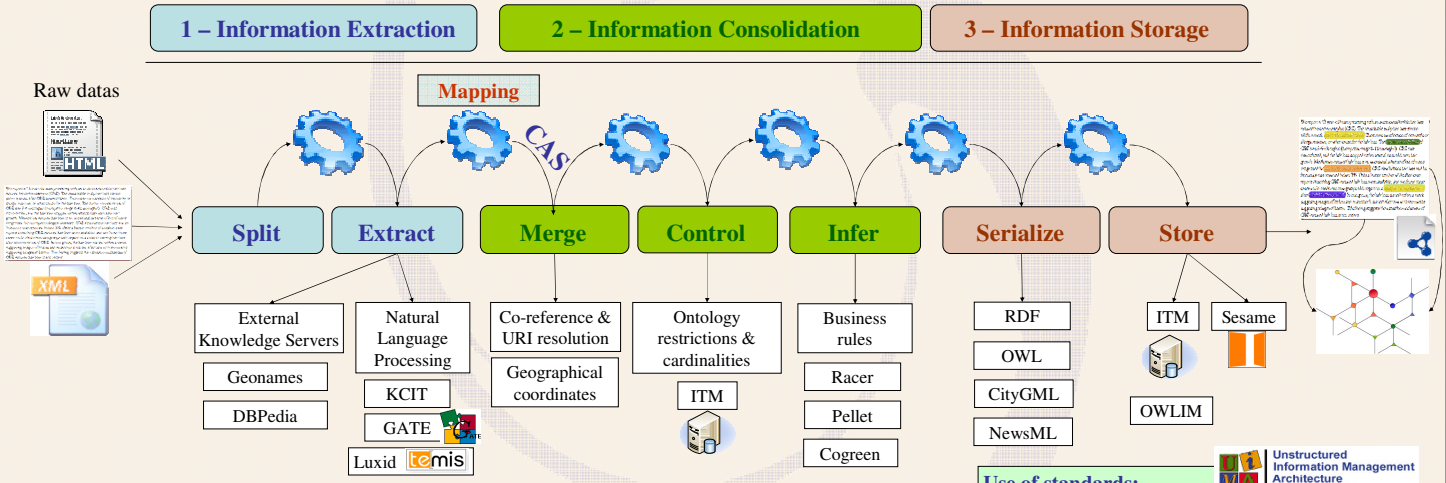
Ontology population:

- Relies on semantic annotation
- The role of human annotators ultimate
- High costs especially for large collections of data

-What is missing in existing tools?

- Compliance with standards
- Flexibility to customize the annotation workflow
- Ontology-based consolidation processes
- Low support for merging automatic and manual processes

Our solution : Content-Augmentation Manager Framework



CAS: Internal data structure, generic for semantic annotation and common within the process – composed of « Entity, Property, Occurrences, Metadata »

Use of standards:

- Orchestration based on the UIMA infrastructure
- Service Oriented Architecture
- Compatible with RDF-OWL-… formats

Evaluation of consolidation algorithms

Methodology:

1) 64 GATE software artefacts annotated (with regards to the GATE ontology) by humans to create gold standard
2) Improving the tool based on comparing results with the gold standard
3) Select another 20 documents to serve as a representative corpus
4) Information extraction evaluation based on KCIT performance

	Precision	Recall
4 Forum posts	98.6111111	100.0
7 chapters of the GATE User Manual	96.0007672	96.9611548
2 Web pages	94.379845	95.0787402
3 publications	89.796798	95.8392268
3 java classes	97.2592593	98.8636364
1 GATE application developers guide	96.484375	98.4063745
Total	94.2830463	96.8763326

5) Information consolidation evaluation

Element type in the ontology	Number of correct elements (A)	Number of missing elements (B)	Number of spurious elements (C)	Recall (A/A+B)	Precision (A/A+C)	F1-measure (R*P)/0.5(R+P)
Kb instances	208	0	64	1	0,765	0,867
Annotations	168	0	12	1	0,933	0,965

Created workflows in existing projects

project	ontology	corpus	CA tool	repository
TAO	PROTON	news article	regular expressions	Sesame
	GATE onto	gate artefacts	KCIT	Sesame
	GATE onto	***gate artefacts***	***KCIT***	***ITM***
Terradata	Architectural ontology (3D objects)	3D objects	DBPedia and Geonames web services	ITM
VigiTermes	Adverse Drug Effect ontology	PubMed abstracts	Luxid (Temis)	ITM
Eiffel	Tourism ontology	Touristic web sites	TimeFrame (Univ Paris X)	ITM
Microbio	MiRNA ontology	PubMed articles	FunGen Discovery (INSERM)	Sesame

- Opensource: http://sourceforge.net/ projects/scan-ca-manager
- Testing interface: http://client2.mmondeca.com/scan