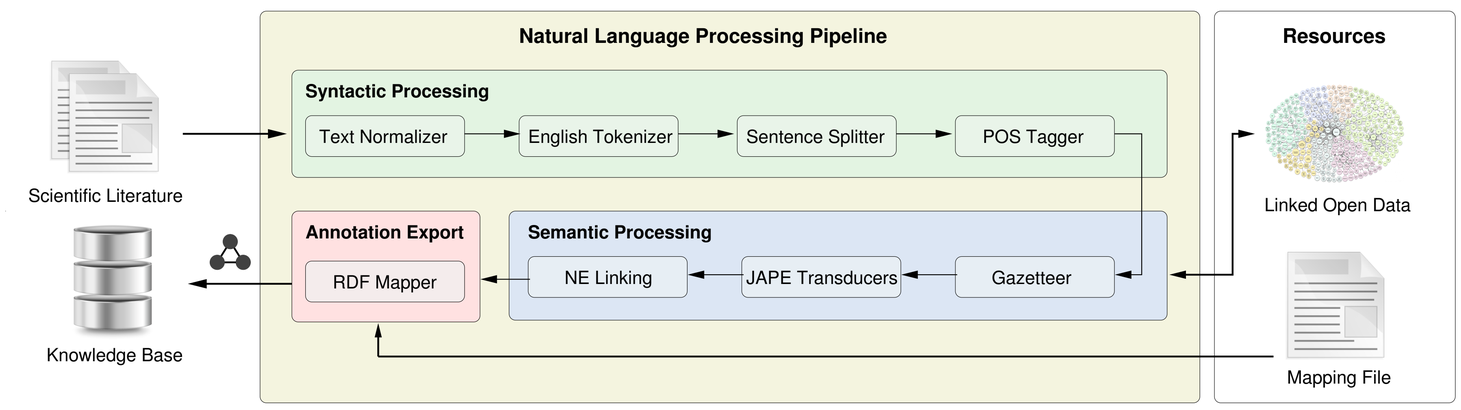
Automatic Construction of a Semantic Knowledge Base from CEUR Workshop Proceedings

Bahar Sateli René Witte

Semantic Software Lab
Department of Computer Science and Software Engineering
Concordia University, Montréal, Canada
{sateli,witte}@semanticsoftware.info

Motivation

- Extract machine-readable, contextual information from scientific literature
- Primarily address Task 2 of the Semantic Publishing Challenge 2015
- Techniques from Natural Language Processing and Semantic Web domains to construct a KB
- Developed a text mining pipeline based on the GATE framework [1]



Syntactic Processing

- Scrape text from input documents and normalization of output
- Break down text into tokens (e.g. words, symbols) and detecting sentence boundaries
- Stemming (i.e., finding root form of tokens) using GATE's Morphological Analyzer
- Annotate tokens with their Part-of-Speech category (e.g., noun, verb, adjective)

Semantic Processing

- Match tokens against Gazetteer lists (dictionaries of known entities like city names)
- Hand-crafted rules for entity detection: *Title, Authors, Affiliations, References, etc.*
- Rules are written in the JAPE language that allows regular expressions over annotations
- Detect *Contributions* and link *Named Entities* in papers to LOD resources [2]

Annotation Export

- We introduce a novel, flexible system to transform annotations to RDF triples
- Designed the PUBlication Ontology (http://lod.semanticsoftware.info/pubo#)
- Triples' semantic types and inter-relationships are identified during runtime
- Transformation process is performed according to a user-provided, custom mapping file
- Mapping file is an RDF document itself and uses Linked Open Vocabularies:

```
### Annotation Mapping ###
map:GATEAuthor a map:Mapping ;
        map:type    foaf:Person ;
        map:GATEtype   "Author" ;
        map:hasMapping map:GATEContentMapping .

map:GATEAffiliation a map:Mapping ;
        map:type    foaf:Organization ;
        map:GATEtype   "Affiliation" ;
        map:hasMapping map:GATEContentMapping ;
        map:hasMapping map:GATELocatedInFeatureMapping .
```

```
### Relation Mapping ###
map:AuthorAffiliationRelationMapping a map:Mapping ;
        map:type rel:employedBy ;
        map:domain map:GATEAuthor ;
        map:range map:GATEAffiliation ;
        GATEattribute "employedBy" .
```

Evaluation

- Comparison against 20 manually annotated papers
- Calculated Precision, Recall and F-1 Measure

Annotation	Prec.B/A	Rec.B/A	F1.0-a.
Abstract_body	0.8333	0.7895	0.8108
Affiliation	0.7857	0.6875	0.7333
Author	0.9437	0.9853	0.9640
Metadata_body	0.9250	0.9250	0.9250
Ref_authors	0.8393	0.9400	0.8868
Ref_source	0.8762	0.8844	0.8803
Ref_title	0.9128	0.8844	0.8984
References_body	0.7250	0.7632	0.7436
Title	1.0000	1.0000	1.0000
Macro summary	0.8712	0.8733	0.8714
Micro summary	0.8762	0.8968	0.8864

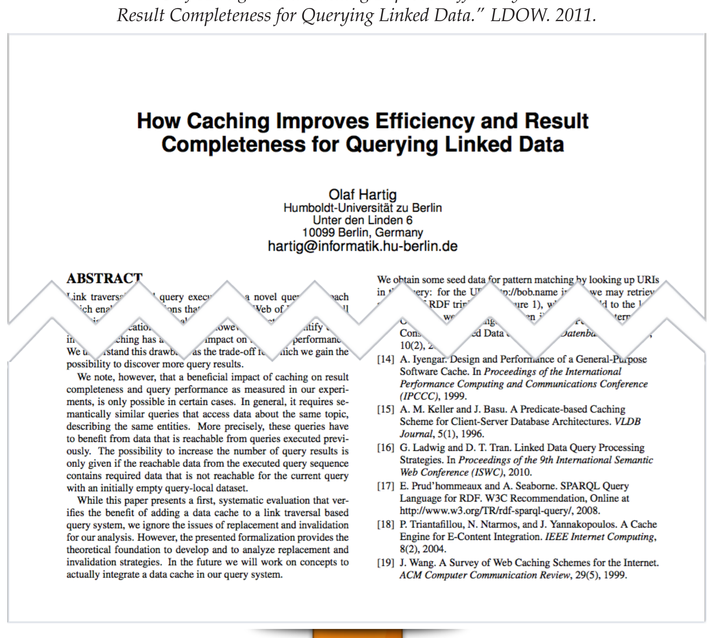
Observations

- Unconventional headers impact segmentation
- Low recall when affiliations were not in English
- Anomalies in bibliographical entries

Running Example

Workflow starts by feeding a paper into the pipeline

Olaf Hartig. "How Caching Improves Efficiency and Result Completeness for Querying Linked Data." LDOW. 2011.

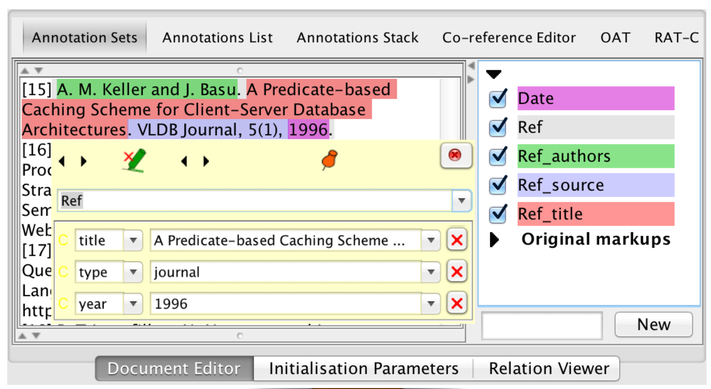

How Caching Improves Efficiency and Result Completeness for Querying Linked Data

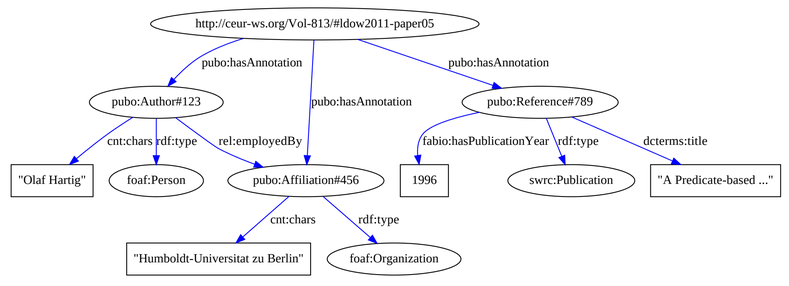
Olaf Hartig
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin, Germany
hartig@informatik.hu-berlin.de

ABSTRACT

```
Rule: reference_authors(
 {Person}
 ({Token.kind=="punctuation",Token.string==","}{Person})*
 (({Token.kind=="punctuation",Token.string==","})?
 {Token.string=="and"} {Person})?
):mention
-->
:mention.Ref_authors  = {debugRule = "reference_authors"}

Rule: reference_title(
 {Ref_authors}
 ({Token.string==":"} | {Token.string=="."})
  (({Token, !Token.string=="."})+)?:title
  {Token.string=="."}
):mention
-->
:title.Ref_title  = {content = :title@cleanString}
```





Example Query

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX pubo: <http://lod.semanticsoftware.info/pubo#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX swrc: <http://swrc.ontoware.org/ontology#>
PREFIX fabio: <http://purl.org/spar/fabio/>

SELECT ?resource_iri ?title ?publication_year WHERE{

 ?paper pubo:hasAnnotation ?resource_iri.
 ?resource_iri rdf:type swrc:Publication.
 ?resource_iri dcterms:title ?title.
 ?resource_iri fabio:hasPublicationYear ?publication_year

 FILTER (?publication_year < 2000)
}
```

resource_iri	title	publication_year
http://ceur-ws.org/Vol-813/#ldow2011-paper05	*"A Predicate-based Caching Scheme for Client-Server Database Architectures."*	1996

References

[1] H. Cunningham et al., **"Text Processing with GATE (Version 6)"**, *University of Sheffield, Department of Computer Science* 2011

[2] B. Sateli, R. Witte, **"What's in this paper? Combining Rhetorical Entities with Linked Open Data for Semantic Literature Querying"**, *SAVE-SD 2015, Florence, Italy. ACM* 2015

Supplementary Materials

Access to SPARQL endpoint and text mining pipeline resources:

http://www.semanticsoftware.info/sempub-challenge-2015

Contact us on Twitter: @SemSoft