# 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

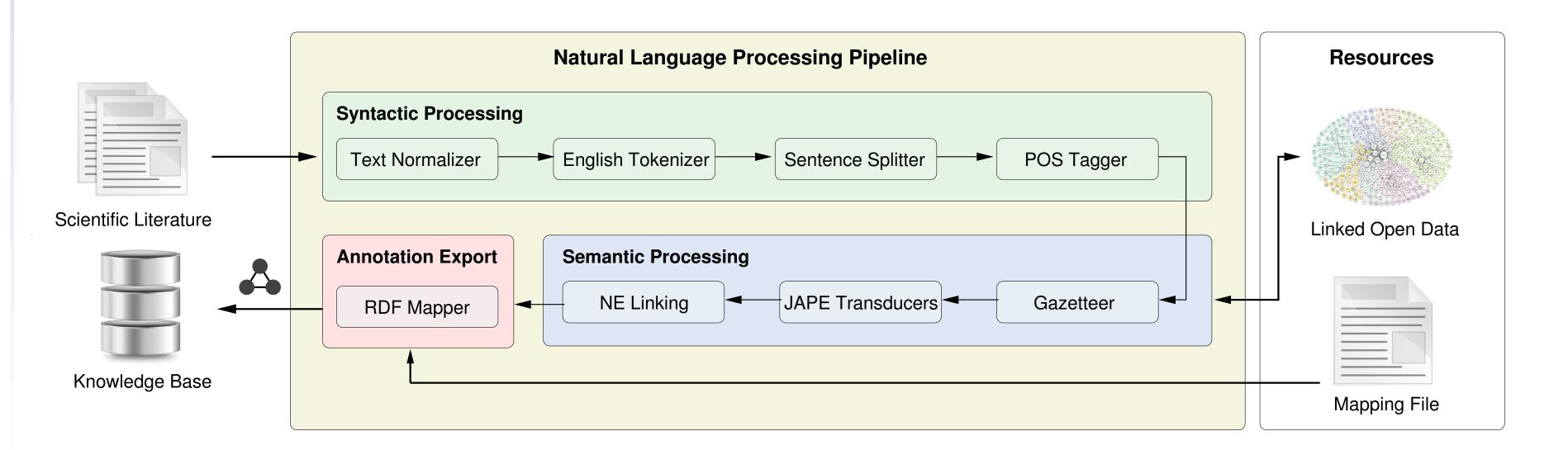
Concordia University, Montréal, Canada

{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 ;
                 foaf:Person ;
       map:type
       map:GATEtype "Author";
       map:hasMapping map:GATEContentMapping .
                                                        ### Relation Mapping ###
                                                        map:AuthorAffiliationRelationMapping a map:Mapping;
map:GATEAffiliation a map:Mapping
                                                               map:type rel:employedBy ;
                 foaf:Organization ;
                                                               map:domain map:GATEAuthor ;
       map:GATEtype "Affiliation";
                                                               map:range map:GATEAffiliation ;
       map:hasMapping map:GATEContentMapping ;
                                                               GATEattribute "employedBy".
       map:hasMapping map:GATELocatedInFeatureMapping .
```

#### Evaluation

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

#### **Observations**

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

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

# 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.

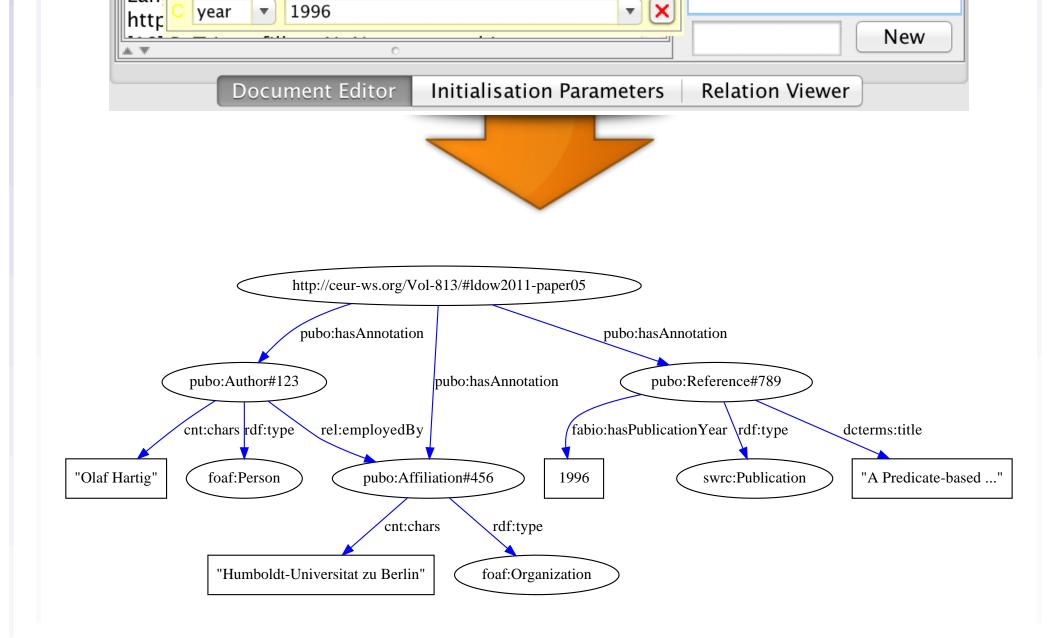
How Caching Improves Efficiency and Result

**Completeness for Querying Linked Data** stand this drawb. as the trade-off to nich we gain the [14] A. Iyengar. Design and Performance of a General-Purpose possibility to discover more query results. Software Cache. In Proceedings of the International We note, however, that a beneficial impact of caching on result Performance Computing and Communications Conference completeness and query performance as measured in our experi-15] A. M. Keller and J. Basu. A Predicate-based Caching mantically similar queries that access data about the same topic, Scheme for Client-Server Database Architectures. VLDB describing the same entities. More precisely, these queries have Journal, 5(1), 1996. to benefit from data that is reachable from queries executed previ-[16] G. Ladwig and D. T. Tran. Linked Data Query Processing ously. The possibility to increase the number of query results is Strategies. In Proceedings of the 9th International Semantic only given if the reachable data from the executed query sequence Web Conference (ISWC), 2010. contains required data that is not reachable for the current query [17] E. Prud'hommeaux and A. Seaborne. SPARQL Query with an initially empty query-local dataset. Language for RDF. W3C Recommendation, Online at While this paper presents a first, systematic evaluation that verhttp://www.w3.org/TR/rdf-sparql-query/, 2008. ifies the benefit of adding a data cache to a link traversal based [18] P. Triantafillou, N. Ntarmos, and J. Yannakopoulos. A Cache query system, we ignore the issues of replacement and invalidation Engine for E-Content Integration. IEEE Internet Computing, for our analysis. However, the presented formalization provides the theoretical foundation to develop and to analyze replacement and invalidation strategies. In the future we will work on concepts to [19] J. Wang. A Survey of Web Caching Schemes for the Internet. actually integrate a data cache in our query system. ACM Computer Communication Review, 29(5), 1999.



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



# Example Query

{Token.string=="."}

):mention

-->

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

resource\_irititlepublication\_yearhttp://ceur-ws.org/Vol-<br/>813/#ldow2011-paper05"A Predicate-based Caching Scheme for<br/>Client-Server Database Architectures."1996

# Supplementary Materials

Access to SPARQL endpoint and text mining pipeline resources: http://www.semanticsoftware.info/sempub-challenge-2015

