ROSeAnn: Taming Online Semantic Annotators[*]

Stefano Ortona, Luying Chen, Giorgio Orsi
Department of Computer Science, Oxford University, Wolfson Building, Parks Road, Oxford OX1 3QD
firstname.lastname@cs.ox.ac.uk

ABSTRACT

Named entity extractors are a popular means for enriching documents with semantic annotations. Both the overlap and the increasing diversity in the capabilities and in the vocabularies of the annotators motivate the need for managing and integrating semantic annotations in a coherent and uniform fashion.

ROSEANN is a framework for the management and the reconciliation of semantic annotations. It provides end-users and programmers with a unified view over the results of multiple online and standalone annotators, linking them to an integrated ontology of their vocabularies, and supporting a variety of document formats such as: plain text, live Web pages, and PDF documents. Although ROSEANN provides two pre-defined algorithms for conflict resolution – one supervised, appropriate when representative training data is available, and one unsupervised – it also allows application developers to define their own integration techniques, as well as extending the pool of annotators as new ones become available.

1. INTRODUCTION

A growing number of resources are available for recognising named entities in documents (e.g. London, the King of Spain) and to link them to particular entity types (e.g. politicians, governmental organizations) generating *semantic annotations*. While originally focused on a limited number of commonly used abstract types such as people, locations, and organizations, the ecosystem of annotators is now increasingly diverse. Modern online annotators support large vocabularies ranging from abstract, e.g., persons, to more specialised entity types, e.g., proteins.

Annotations play an important role, e.g., in semantic search engines, information extraction, and for the automated production of linked data. Many annotation services are nowadays freely available online (e.g. OPENCALAIS, ZEMANTA). This gives the potential for developers to easily embed annotation-based functionalities in their applications. But doing this leads to many challenges, including dealing with multiple formats, judging the quality of annotations and reconciling disagreeing opinions about an entity.

[*]The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013) / ERC grant agreement DIADEM, no. 246858.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2578038.

Consider the example in Figure 1. Here we see the variations in quality within annotators, as well as an idea of several flavors of clash in annotator opinions. The token "*Japanese*" is labeled as an *Organisation* by WIKIMETA, as a *Language* by DBPEDIA SPOTLIGHT, and as a *Country* by EXTRACTIV– clearly these outputs are incompatible, since these three entity types have mutually disjoint meaning in a given context.

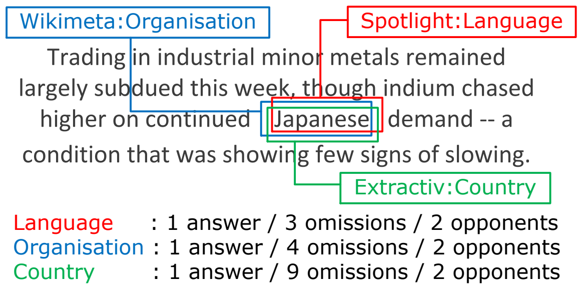

Wikimeta:Organisation
Spotlight:Language
Trading in industrial minor metals remained largely subdued this week, though indium chased higher on continued Japanese demand -- a condition that was showing few signs of slowing.
Extractiv:Country
Language : 1 answer / 3 omissions / 2 opponents
Organisation : 1 answer / 4 omissions / 2 opponents
Country : 1 answer / 9 omissions / 2 opponents

Figure 1: Conflicting and re-enforcing annotations.

We will now provide a user- and programmer-oriented overview of ROSEANN, a framework for manipulation, querying and reconciliation of semantic annotations provided by multiple independent annotators. ROSEANN concurrently annotates text or Web documents with multiple independent annotators, and then presents a unified view as a single annotated document.

ROSEANN allows the invocation of online annotation services, retrieval and manipulation of annotations, conflict analysis, and exporting the annotated documents as XML. Most importantly, it allows for *aggregation* of annotations, returning a logically-consistent set of annotations w.r.t. a background ontology. ROSEANN supports a variety of document formats such as plain-text, live Web pages, and PDF documents.

Our prior work [1] has shown that ROSEANN's aggregated annotations are of significantly higher quality, compared to those produced by either individual annotators or other reconciliation techniques, e.g., [7].

ROSEANN currently supports eleven annotators, namely: OPENCALAIS, EXTRACTIV, DBPEDIA SPOTLIGHT, ALCHEMYAPI, ZEMANTA, LUPEDIA, WIKIMETA, SAPLO, YAHOOYQL, STANFORDNER, and NETAGGER, as well as custom annotators based on the GATE framework [2] for specific domains.

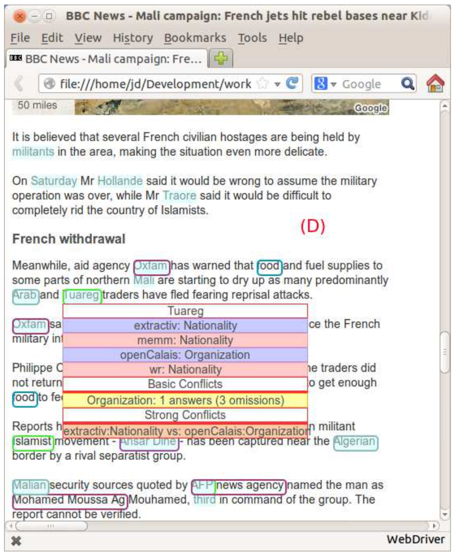
ROSEANN also supports two aggregation methods: one *supervised* and one *unsupervised* and uses the reconciled entity types to give links to external LOD resources.

Additional annotators and reconciliation algorithms can be seamlessly added by means of the ROSEANN API.

2. ROSEANN GUI

The easiest way to use ROSEANN's semantic annotation and reconciliation capabilities is via its standalone graphical interface. The GUI supports the annotation of **(i)** plain text (Figure 3), **(i)** PDF (Figure 4), and **(i)** live Web documents (Figure 2).

Web navigation is provided by driving a Firefox web browser via Selenium WebDriver[1]. PDF processing is based on the Apache PDFBox Library[2].



BBC News - Mali campaign: French jets hit rebel bases near Kid
File Edit View History Bookmarks Tools Help
BBC News - Mali campaign: Fre...
file:///home/jd/Development/work
Google
50 miles
(D)
WebDriver

It is believed that several French civilian hostages are being held by militants in the area, making the situation even more delicate.

On Saturday Mr Hollande said it would be wrong to assume the military operation was over, while Mr Traore said it would be difficult to completely rid the country of Islamists.

French withdrawal

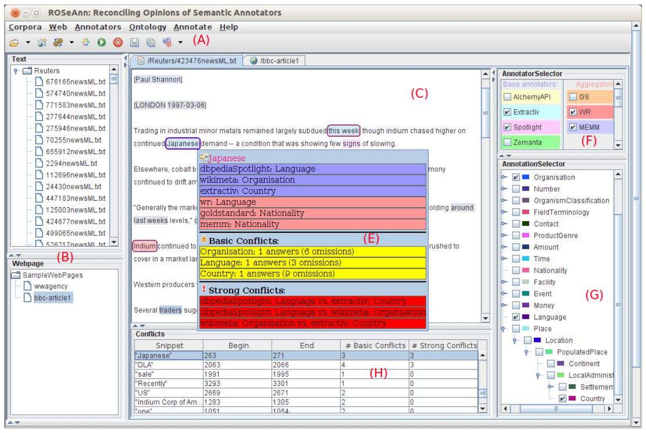
Meanwhile, aid agency Oxfam has warned that food and fuel supplies to some parts of northern Mali are starting to dry up as many predominantly Arab and Tuareg traders have fled fearing reprisal attacks.

Tuareg
extractiv: Nationality
memm: Nationality
openCalais: Organization
wr: Nationality
Basic Conflicts
Organization: 1 answers (3 omissions)
Strong Conflicts
extractiv:Nationality vs. openCalais:Organization

Reports h... militant Islamist movement - Ansar Dine - has been captured near the Algerian border by a rival separatist group.

Malian security sources quoted by AFP news agency named the man as Mohamed Moussa Ag Mouhamed, third in command of the group. The report cannot be verified.

Figure 2: Web Annotation.



Japanese
dbpediaSpotlight: Language
wikimeta: Organisation
extractiv: Country
wr: Language
goldstandard: Nationality
memm: Nationality
Basic Conflicts:
Organisation: 1 answers (6 omissions)
Language: 1 answers (3 omissions)
Country: 1 answers (9 omissions)
Strong Conflicts:
dbpediaSpotlight: Language vs. extractiv: Country
dbpediaSpotlight: Language vs. wikimeta: Organisation
wikimeta: Organisation vs. extractiv: Country

Snippet	Begin	End	# Basic Conflicts	# Strong Conflicts
"Japanese"	263	271	3	3
"DLA"	2063	2066	4	3
"sale"	1991	1995	1	0
"Recently"	3293	3301	1	0
"US"	2669	2671	2	0
"Indium Corp of Am...	1283	1305	2	0

Figure 3: Text Annotation.

The main menu bar on the top of the GUI (**A**) supports document loading, annotation, ROSEANN configuration, as well as the possibility to save the annotated documents and to browse the entity types of the mapping ontology via the ROSEANN SPARQL Endpoint[3]. Annotated documents are accessible from the left-hand side of the GUI (**B**).

Textual documents are visualized in the main text area of the tool (**C**), while web documents are loaded in an independent browser

[1] http://docs.seleniumhq.org/.

[2] http://pdfbox.apache.org.

[3] http://163.1.88.38:8081/openrdf-workbench/.

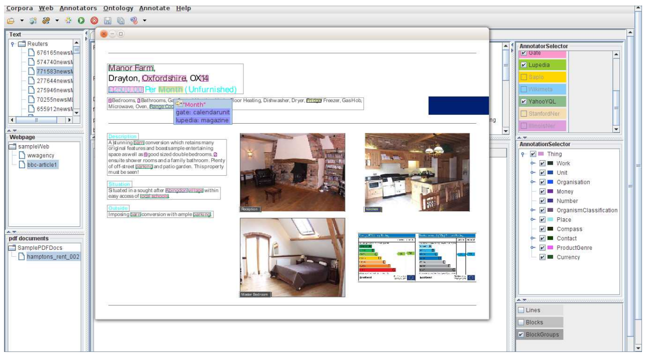


Figure 4: PDF Annotation.

window (**D**). In both cases, the user can interact with the document and the web browser before starting the annotation process.

After a document has been annotated, ROSEANN highlights the recognized entities in the main text area or in the web browser. The highlighting consists of a colored border around the identified entities representing an entity type in the ontology, while a different background color represents the annotator or aggregator recognizing that particular entity.

By hovering over the highlighted entities, ROSEANN provides the list of opinions for all annotators and aggregators specifying which annotator contributed a given entity type, together with the conflicts (**E**). A different background on the tooltip is used to distinguish the opinions of annotators from those of the aggregators. When an annotator provides also links to LOD, e.g., to DBPedia, ROSEANN makes available those anchors to the user in the tooltip.

On the right-hand-side of the GUI we list all annotators (**F**) that identified at least one entity in the current document and the identified entity types organized into a hierarchy (**G**) that corresponds to the structure of our mapping ontology. The user can decide which annotators, aggregators and entity types to visualize in the main text-area or in the browser.

At the bottom of the GUI we provide a table listing all conflicts generated by the annotators in the given document (**H**). In particular, we report (from the left to the right in the table) the text snippet involved in the conflict, the start and end offset of the text span in the document, and the number and type of conflicts occurring on that span. After selecting a row in the conflict table, ROSEANN blinks the span involved in the conflicts in the text area or in the browser.

Annotated documents can be exported in a variety of formats such as XML, CSV, and SQL for import into a relational database.

3. ROSEANN API

ROSEANN comes with a Java API and an online REST-ful service. The main abstractions underlying the ROSEANN API are:

(1) Annotators. This is the basic abstraction in ROSEANN and it represents a standalone or an online annotator together with its configuration parameters. Application developers can instantiate and configure an annotator using parameters such as the type of concepts to return and a timeout, as well as specific parameters of the annotator when known. For those annotators requiring a user-key, ROSEANN allows application developers to add their private keys for use with ROSEANN.

(2) Document models. An abstraction for representing plain-text, PDF, and HTML documents in a uniform fashion.

(3) Annotated document model. This abstraction represents an annotated document model. The API supports retrieval of annotations by entity types (e.g., *Person*), annotators (e.g., OPENCALAIS) and by document spans, e.g., a [start, end] character range within the document. Also, application developers can retrieve conflicts by type (e.g., logical and omission), and by document span.

(4) Individual annotators can be combined into an annotation pool, an abstraction used to invoke multiple annotators in parallel on a given set of resources (i.e., documents). The advantage of annotation pools is the possibility to share configuration parameters that are common to multiple annotations, thus avoiding the burden of configuring each annotator in the same way. When dealing with HTML documents, ROSEANN supports the annotation of different sections of the DOM with different annotation pools, e.g., DOM *elements*, *attributes*, and *scripts*. This is particularly useful in certain applications where it is important to distinguish annotations on visible parts of the DOM (e.g., for information extraction), from those on invisible content such as attributes and scripts (e.g., for program analysis). An example of annotation-pool configuration is shown in Figure 5 where different annotators are used for different sections of the DOM.

(5) The reconciliation of sets of annotations coming from different annotators is carried out via an *aggregator*. This abstraction represents a reconciliation algorithms, e.g., MEMM (Maximum Entropy Markov Models) [1]. Application developers can extend ROSEANN with their own aggregation algorithms and then use the ROSEANN API and GUI to manipulate and display the results. A particular class of aggregators is the *trainable* aggregator, representing aggregators algorithms requiring training. ROSEANN provides abstractions to carry out training of trainable aggregators programmatically.

```
<domain>
  <domannotator>
    <elements>
      <annotationpool>
        <annotator>opencalais</annotator>
        <timeout>10000</timeout>
      </annotationpool>
    </elements>
    <attributes>
      <annotationpool>
        <annotator>AlchemyAPI</annotator>
        <annotator>Lupedia</annotator>
        <timeout>15000</timeout>
      </annotationpool>
    </attributes>
    <properties>
      <annotationpool>
        <annotator>Wikimeta</annotator>
        <annotator>opencalais</annotator>
        <timeout>150000</timeout>
      </annotationpool>
    </properties>
  </domannotator>
</domannotator>
```

Figure 5: Example configuration for DOM annotation.

4. APPLICATIONS OF ROSEANN

Apart from the natural application of ROSEANN in entity extraction systems, ROSEANN is currently employed in several projects, ranging from security to structured web data extraction, and in which it has proved its versatility.

The main use of ROSEANN is DIADEM [4], a project at Oxford University on unsupervised domain-specific web object extraction. Its goal is to transform unstructured web information into highly structured data without human supervision. DIADEM replaces human annotators in traditional wrapper induction systems with ROSEANN, coupling it with knowledge about the application domain and about the representation of these entities in the textual, structural, and visual features of a website of this domain. With this knowledge at hand, DIADEM needs to analyze only a small fraction of a web site to automatically generate (induce) a wrapper that is then executed in the second stage to extract all the relevant data on the site. In this setting the ability of ROSEANN to integrate multiple annotators is of paramount importance to reduce the need for domain-specific annotators. Also, the ability of ROSEANN to annotate different sections of the DOM with different annotation pools, together with its reconciliation capabilities, reduce the noise in the annotations that is the main source of errors in annotation-driven wrapper inducers such as [3]. Figure 6 shows the use of

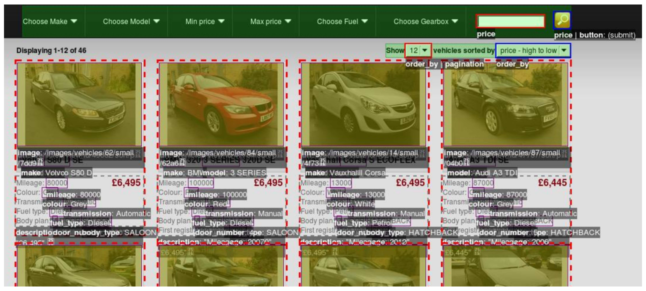


Figure 6: ROSeAnn within DIADEM.

ROSEANN within DIADEM, in particular, for the unsupervised segmentation of classified listings on the web [6] and understanding of forms [5].

5. REFERENCES

[1] L. Chen, S. Ortona, G. Orsi, and M. Benedikt. Aggregating semantic annotators. *PVLDB*, 6(13):1486–1497, 2013.

[2] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters. *Text Processing with GATE (Version 6)*. 2011.

[3] N. N. Dalvi, R. Kumar, and M. A. Soliman. Automatic wrappers for large scale web extraction. *PVLDB*, 4(4):219–230, 2011.

[4] T. Furche, G. Gottlob, G. Grasso, O. Gunes, X. Guo, A. Kravchenko, G. Orsi, C. Schallhart, A. J. Sellers, and C. Wang. Diadem: domain-centric, intelligent, automated data extraction methodology. In *Proc. of WWW*, pages 267–270, 2012.

[5] T. Furche, G. Gottlob, G. Grasso, X. Guo, G. Orsi, and C. Schallhart. The ontological key: Automatically understanding and integrating forms to access the deep web. *The VLDB Journal*, 22(5):615–640, 2013.

[6] T. Furche, G. Gottlob, G. Grasso, G. Orsi, C. Schallhart, and C. Wang. Little knowledge rules the web: Domain-centric result page extraction. In *Proc. of RR*, pages 61–76, 2011.

[7] G. Rizzo and R. Troncy. Nerd: A framework for unifying named entity recognition and disambiguation extraction tools. In *Proc. of EACL*, pages 73–76, 2012.