An Architecture for Data and Knowledge Acquisition for the Semantic Web: the AGROVOC Use Case

Maria Teresa Pazienza, Armando Stellato, Alexandra Gabriela Tudorache, Andrea Turbati, and Flaminia Vagnoni

University of Rome Tor Vergata
Via del Politecnico 1, 00133 Rome, Italy

{pazienza, stellato, tudorache, turbati}@info.uniroma2.it
f.vagnoni@gmail.com

Abstract. We are surrounded by ever growing volumes of unstructured and weakly-structured information, and for a human being, domain expert or not, it is nearly impossible to read, understand and categorize such information in a fair amount of time. Moreover, different user categories have different expectations: final users need easy-to-use tools and services for specific tasks, knowledge engineers require robust tools for knowledge acquisition, knowledge categorization and semantic resources development, while semantic applications developers demand for flexible frameworks for fast and easy, standardized development of complex applications. This work represents an experience report on the use of the CODA framework for rapid prototyping and deployment of knowledge acquisition systems for RDF. The system integrates independent NLP tools and custom libraries complying with UIMA standards. For our experiment a document set has been processed to populate the AGROVOC thesaurus with two new relationships.

1 Introduction

Nowadays we are surrounded by huge amounts of information: The World Wide Web, books, business documents, all creating a nearly infinite source of unstructured or weakly structured data. It is becoming mandatory to find novel ways to access, categorize and, most important, to extract meaningful knowledge from this virtual repository. Furthermore, to efficiently exploit such knowledge, it is necessary to structure it in a computer readable format. Processing, transforming and manipulating heterogeneous information is not an easy task and requires the integration of several dedicated tools. Frameworks for the integration, orchestration and harmonization of the different aspects related to content acquisition are thus an emerging need for industry-standard knowledge management systems.

In this work, we propose a knowledge elicitation scenario, where we report on the adoption of the CODA[1] framework [1] in a complete content production process,

[1] CODA – Computer Aided Ontology Development, http://art.uniroma2.it/coda

ranging from content analytics, information extraction and triplification for Ontology development and enrichment.

The system aims to extract relations from unstructured web and textual documents to populate the AGROVOC[1] thesaurus with new relationships.

Currently the AGROVOC thesaurus contains more than 40 000 concepts in 21 languages and is widely populated with a few general semantic relations, such as: *broader term*, *narrower term*, *is used as*, *is part of*. Then, there are a series of more specific relations which have been only partially instantiated or even just defined, for future use. For a human expert, identifying such relations in free text content, with a good degree of confidence, requires not only specific skills but also a great amount of time. In the context of this experience, two different relations (and their inverses) will be extracted: *IsPestOf* and *IsInsecticideFor*. The *IsPestOf* relation is defined, but still not instantiated in AGROVOC, while *IsInsecticideFor* was not defined. The closest relation to *IsInsecticideFor* currently defined is *is_use_of* (e.g. "pesticide" <is use of> "ddt"), that relates one substance to its use as an insecticide, but not to its target pests.

This paper presents some notes on related work, the system architecture, the implementation choices and the experimental results. Moreover, conclusions will be drawn and future work will be discussed.

2 Related Work

In recent years, the research community showed a growing interest in the area of relation extraction and semantic role labeling, with different approaches being formulated as Machine Learning or Rule Based Systems [2]. While, machine learning systems have a better generalization power, they typically require extensive training and at least some seed data annotated by domain experts, rule based systems are better suited for specific tasks, but need a comprehensive support of domain experts. Considering that our application is domain related (Agriculture), and that we aim to perform high precision Relation Extraction for enriching AGROVOC (one of the biggest semantic resources on the Agriculture field), the second approach was selected. While maintaining an adequate recall, in the presented scenario, the precision was very important, to minimize the human experts' effort for validating the extracted triples.

Recently, several rule based information extraction systems were developed. Among them we cite: TextMarker [3], the AVATAR Information Extraction System [4] and, different systems for the medical [5] and biological (gene analysis) domains [6]. Furthermore, in the last decade, a number of ontology development related research directions emerged as: the automation of ontology development (KYOTO Project [7]), ontology and lexicon integration [8, 9], or ontology learning and population [1, 10]. In this context, open platforms and frameworks as GATE [11] and UIMA [12] were developed. Moreover, business oriented systems were designed as the ones for AGROVOC thesaurus enrichment including tasks as automatic term relationship cleaning and refinement [13] and vocabulary alignment [14].

[1] AGROVOC – FAO's Agricultural Thesaurus, www.fao.org/agrovoc/

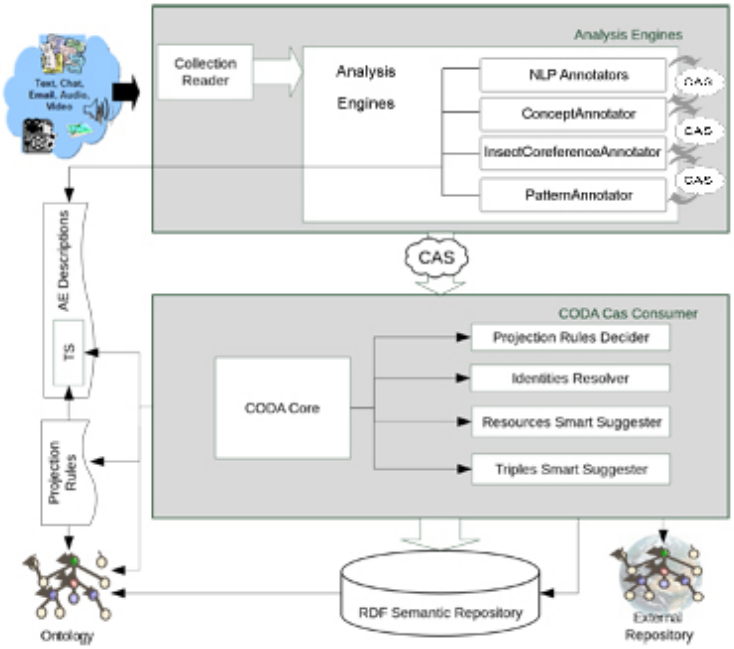

Fig. 1. System Architecture

While preexistent systems focus on general relations (broader and narrower term), the proposed system focuses on specific semantic relations extraction from free text, and concept alignment between different resources (e.g. *IsPestOf, IsInsecticideFor*).

3 Architecture

The architecture aims to optimize the Information Acquisition flow and to support the development of an easy and fast to integrate, modular system, complying with UIMA standards (see **Fig. 1**). The system is based on CODA framework and combines different NLP tools together with custom libraries and implements a complete NLP flow starting with collection reader, Information Extraction, more specifically relation extraction and ending with triplification and ontology enrichment/development.

In this context, the system implements different Analysis Engines, ranging from the reuse of available NLP annotators based on Stanford NLP Suite[1] and a few dedicated ones such as: *ConceptAnnotator*, *InsectCoreferenceAnnotator* and *PatternAnnotator*. Each annotator adds one ore more layers of semantic information having different roles in the workflow. NLP Annotators include: Segmenter, PosTagger, Lemmatizer, NamedEntityRecognizer, Parser and CoreferenceResolver [15]. ConceptAnnota-

[1] Stanford Core NLP - http://nlp.stanford.edu/software/corenlp.shtml

Table 1. Triplification process

The sentence: "In the eastern US, the **gypsy moth*[c_30232]*** prefers **oaks*[c_6409]***, **poplar*[c_541]***, **apple*[c_6116]***…" (Excerpt form testing file no 136).

Becomes the following **RDF triples:**

```
<rdf:Description rdf:about="http://aims.fao.org/aos/agrovoc#c_30232">
   <rdf:type rdf:resource="http://aims.fao.org/aos/agrovoc#Insect"/>
   <isPestOf xmlns="http://aims.fao.org/aos/agrovoc#"
             rdf:resource="http://aims.fao.org/aos/agrovoc#c_6409"/>
   <isPestOf xmlns="http://aims.fao.org/aos/agrovoc#"
               rdf:resource="http://aims.fao.org/aos/agrovoc#c_541"/>
   <isPestOf xmlns="http://aims.fao.org/aos/agrovoc#"
             rdf:resource="http://aims.fao.org/aos/agrovoc#c_6116"/>
</rdf:Description>
```

***Where: c_XXXX** stands for the AGROVOC concept ID*

tor's role is to annotate Insects and Plants found in text using existing resources as AGROVOC and NAL[1] thesauri. Concepts from other resources are aligned with AGROVOC concepts and concepts not present in AGROVOC are suggested as candidates for AGROVOC enrichment. InsectCoreferenceAnnotator is an ad-hoc coreference resolver that recognizes and tags different biological development forms of previously annotated insects as: eggs, larvae, pupae, and PatternAnnotator is an extensible annotator that recognizes different relations present in the analyzed text, each pattern being implemented as a different java class. The complete list of relations and correspondent patterns is presented in **Section 7**.

The role of building the Semantic Repository starting from the UIMA annotated data was delegated to CODA framework. For every annotated document, several RDF[2] triples are generated. Each triple represents a relation having one Subject and one Object. In the case of multiple subjects and/or objects, several triples will be generated one for each subject/object pair. Furthermore, the triples representing the inverse relations will be added (see **Table 1**).

4 Implementation

To start, a local domain knowledge base was constructed, by importing, merging and aligning the concepts of interest for the specific task. This newly created resource is based on AGROVOC and NAL thesauri. To include different forms of concepts and verbs a lemmatizer was used. Moreover, the lists of verbs expressing relations were expanded with their synonyms found in WordNet [16].

To extract *IsPestOf* and *IsInsecticideFor* relations eight different patterns were developed as summarized in **Table 2Error! Reference source not found.**. For each relation the aim was to use generalized rules, to avoid overfitting. Furthermore, given

[1] NAL, National Agricultural Library's Agricultural Thesaurus, http://agclass.nal.usda.gov
[2] RDF, Resource Description Framework, http://www.w3.org/TR/rdf-primer

Table 2. Relation Extraction Pattern Summary

Relation	Pattern	Notes
IsPestOf	(insect)+ (MD)? be (pest|pests) (plant)+	This pattern recognizes if a previously annotated *insect(s) (be) pest of plant(s).*
	(insect)+ (VB (including synonyms)) (plant)+	VB in {attack, damage, devastate, eat, feed on, feed upon, prefer, munch, enter, crawl, tunnel, infest, destroy} and synonyms.
IsInsecti-cideFor	(pesticide)+ (MD)? be (pesti-cide(s)|insecticide(s)) (insect)+	One or more general pesticides (chemicals) are identified as insecticide(s) for specific insect(s).
	(pesticide)+ (MD)? be (used| (effec-tive (including synonyms)) (against|on) (insect)+	[effective] ADJ and its synonyms.
	(pesticide)+ (VB (including syno-nyms)) (insect)+	VB in {repel, control, destroy, mitigate} and synonyms VB in {kill, defeat} without synonyms.
	(insect)+ (MD)? be (sensitive to|sprayed with) (pesticide)+	Identifies if insect(s) is/are sensitive to specific chemical.
General Patterns	Interrogative Pattern	Used to eliminate interrogative phrases that can introduce ambiguity
	Negation Pattern	Used to eliminate negative phrases that can intro-duce ambiguity

Where: MD – modal verb (can, may, could, should ecc.); VB - verb

the scope of this demo, we selected a small number of rules to speed up the annotation, development, experimental process and evaluation.

5 Corpus description

The analyzed corpus has been collected as a selection of documents from a few publicly available websites (e.g. wikipedia.org, usda.gov, agric.wa.gov.au). This added different degrees of complexity to the analysis, due to the presence of not correctly formatted phrases (implicit subject or verb) or very long and complex ones, difficult to be understood even for human experts. For instance, in many cases the verb is missing as in: *"Moths of economic significance:* Gypsy moth *(Lymantria dispar), a pest of hardwood trees in North America."*.

Furthermore, often coreference analysis could not be applied as in several sentences the subject referred to the concept from the precedent section or in other cases was a specific instance of the concept present in the preceding sentence. *E.g. "The most serious pests are mealybugs that feed on citrus;* ***other species*** *damage sugarcane, (…)"*. It is difficult even for a human to understand that "other species" refers to other mealybugs not eating citrus.

The entire corpus contains 270 HTML files, of different lengths ranging from few sentences to several pages, each covering a specific arguments as insect pests (75%) or pesticides (25%), and was divided into training (170 files) and testing (100 files).

6 Experimental Setup and Evaluation Criteria

The aim of the experiments was to evaluate the relation extraction flow, including different annotators' contribution for two different relations: *IsPestOf* and *IsInsecticideFor*. In this context, different experiments were conducted, analyzing both HTML and well formatted txt files. Txt files were extracted from HTML pages and cleaned (tags, links and other specific HTML information was discarded).

Furthermore, as several sentences are difficult to be understood even by domain experts, both the existing resources and sentence complexity were considered for evaluation. Sentences in which a clear coreference relation was not found, or implied concepts and relations are present (subject or verb are missing), were not accounted (about 30%, mostly regarding insect pests). E.g. *"In late 2007, the moth eradication program involving (aerial spraying) of a product containing E. postvittana attractant sex pheromones as its active ingredient, among other substances not yet revealed to the public, over sixty square miles near the Pacific coast between Monterey and Santa Cruz was begun."*. In this case sex pheromones are not an insecticide but they can be used in an integrated pest control program.

To understand the contribution of each general annotator we designed four experiments over the baseline.

Baseline experiment includes ConceptAnnotator and PatternAnnotator without synonyms and InsectCoreferenceAnnotator. Furthermore, interrogative and negative sentences were not excluded and this is expected to introduce errors. Negative sentences are the sentences containing negation words as *not* and several negative adverbs and their synonyms such as: *never, uncommon, inconceivable, out of the question, unimaginable, unacceptable*. ***Experiment1*** consists on the baseline experiment to which was added the InsectCoreferenceAnnotator. ***Experiment2*** is the same as *Experiment1* from which interrogative and negative sentences were excluded. ***Experiment3*** adds synonyms to *Experiment2, while* ***Experiment4*** adds synonyms to *Experiment1*. In the context of AGROVOC enrichment, for each experiment the number of distinct extracted triples was analyzed. Results were evaluated using precision and recall as in Information Retrieval (see **Section 7*).***

7 Results

The objective of preferring precision over recall was achieved as shown in **Table 3**. Each annotator has its own positive or negative contribution to the results. The best results were obtained within ***Experiment4***, while ***Experiment1*** had a major contribution to results improvement.

Table 3. Experimental results

	HTML Corpus								
Relation	***IsPestOf***			***IsInsecticideFor***			***AllRelations***		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
Baseline	***82.4***	***24.3***	***37.6***	***100.0***	***27.9***	***43.6***	***87.0***	***24.7***	***38.5***
Exp.1	87.2	34.5	49.4	100.0	30.2	46.4	90.0	33.3	48.6
Exp.2	87.2	28.6	43.0	100.0	25.6	40.7	90.0	27.8	42.5
Exp.3	83.3	25.2	38.7	100.0	25.6	40.7	87.2	25.3	39.2
Exp.4	**86.3**	**37.0**	**51.8**	**100.0**	**30.2**	**46.4**	**89.1**	**35.2**	**50.4**
	Txt Corpus								
Relation	***IsPestOf***			***IsInsecticideFor***			***AllRelations***		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
Baseline	***75.0***	***14.5***	***24.3***	***100.0***	***20.9***	***34.6***	***81.8***	***16.2***	***27.0***
Exp.1	90.6	38.7	54.2	100.0	23.3	37.7	92.1	34.7	50.4
Exp.2	90.2	37.1	52.6	100.0	23.3	37.7	91.8	33.5	39.1
Exp.3	91.1	41.1	56.7	100.0	30.2	46.4	92.8	38.3	54.2
Exp.4	**91.4**	**42.7**	**58.2**	**100.0**	**30.2**	**46.4**	**93.0**	**39.5**	**55.5**

Furthermore, even if the hypothesis of eliminating Interrogative and Negative sentences (***Experiment2*** and ***Experiment3***) seemed promising in theory, the results showed that relevant information is lost.

Text files showed a higher precision as a result of well formatted corpus and *IsInsecticideFor* high precision was achieved being described by a simpler language.

8 Conclusions and Future Work

This paper presented a real life use case. The goal was to extract relations such as: *IsPestOf* and *IsInsecticideFor* from unstructured web documents for enriching FAO's AGROVOC thesaurus. This experience showed that it is possible to implement a fully functional Ontology Development workflow based on CODA architecture including different UIMA Annotators both for undertaking Information Extraction specific tasks and as wrappers to preexisting NLP tools. The developed system showed promising results in terms of thesaurus alignment, relation extraction and ontology population. The flexibility and modularity of the system permit both to expand the current analyzed relations and to change with little effort the application domain, by redefining the information sources and the extraction rules.

To contrast the effort needed for rule generation the system could be refined by learning new rules as in [17]. In this context, other relations as ***[Plant X]*** *benefits from* ***[Substance Y]*** or ***[Substance X]*** *is herbicide for* ***[PestPlant Y]*** could be analyzed and an ad-hoc coreference resolution system for insecticides could further improve results.

Moreover, to avoid creating domain related annotators the pattern matching algorithms could be rewritten in PEARL language [18] under CODA. And the syntactic analysis could be improved and internationalized by integrating different parsers as Chaos [19] that analyzes both English and Italian languages.

References

1. Fiorelli, M., Pazienza, M.T., Petruzza, S., Stellato, A, Turbati, A.: Computer-aided Ontology Development: an integrated environment. New Challenges for NLP Frameworks, Valletta, Malta, May, 18, (2010)
2. Chang, Chia-Hui, Kayed, M., Girgis, M. R., Shaalan, K. F.: A Survey of Web Information Extraction Systems. IEEE Transactions on KDE, pp. 1411-1428, October, (2006)
3. Kluegl, P., Atzmueller, M., Puppe, F.: TextMarker: A Tool for Rule-Based Information Extraction. In: Unstructured Information Management Architecture (UIMA), 2nd UIMA@GSCL Workshop, 2009 Conference of the GSCL (2009)
4. Jayram, T. S., Krishnamurthy, R., Raghavan, S., Vaithyanathan, S, Zhu, H.: Avatar Information Extraction System. In: IEEE Data Eng. Bull., pp. 40-48, IEEE (2006)
5. Regev, Y., et al.: Rule-based extraction of experimental evidence in the biomedical domain: the KDD Cup 2002 (task 1). In: SIGKDD vol. 4 (2), pp. 90-92, ACM (2002)
6. Mykowiecka, A., Marciniak, M., Kupsc, A.: Rule-based information extraction from patients' clinical data. In: Journal of Biomedical Informatics 42(5): 923-936 (2009)
7. Vossen P., Soroa A., Zapirain B., Rigau G.: Cross-lingual event-mining using wordnet as a shared knowledge interface. In Proceedings of GWC'12, Japan. January (2012)
8. Pazienza, M. T., Stellato, A.: Linguistic Enrichment of Ontologies: a methodological framework. In: OntoLex2006, Genoa, Italy (2006)
9. Buitelaar P, Cimiano P, Haase P, Sintek M.: Towards Linguistically Grounded Ontologies. Paper presented at: In Proceedings of ESWC2009 (2009).
10. Cimiano, P.: Ontology Learning and Population from Text Algorithms, Evaluation and Applications XXVIII. Springer (2006)
11. Cunningham, H.: GATE, a General Architecture for Text Engineering. Computers and the Humanities; 36: 223-254 (2002)
12. Ferrucci D, Lally A.: Uima: an architectural approach to unstructured information processing in the corporate research environment. Nat. Lang. Eng. 2004; 10(3-4): 327-348.
13. Morshed, A., Keizer, J., Johannsen, G., Stellato, A., Baker, T.: From AGROVOC OWL Model towards AGROVOC SKOS Model. FAOAIMS (2010)
14. Morshed, A., Sini, M.: Creating and aligning controlled vocabularies. Report. (2009)
15. Lee, H. et al.: Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In: CoNLL-2011 Shared Task (2011)
16. Miller, G. A.: WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41 (1995)
17. Liu, Bin, Chiticariu, L., Chu, V., Jagadish, H. V., Reiss F.: Automatic Rule Refinement for Information Extraction. PVLDB 3(1): 588-597 (2010)
18. Pazienza, M.T., Stellato, A., Turbati, A.: PEARL: ProjEction of Annotations Rule Language, a Language for Projecting (UIMA) Annotations over RDF Knowledge Bases, In: International Conference on Language Resources and Evaluation (LREC 2012) Istanbul, Turkey, May 21-27 (2012)
19. Basili, R., Zanzotto, F. M.: Parsing Engineering and Empirical Robustness. In: Journal of Natural Language Engineering, 8/2-3 June (2002).