Squeezing lemon with GATE

Brian Davis, Fadi Badra, Paul Buitelaar,
Tobias Wunner and Siegfried Handschuh,

Digital Enterprise Research Institute,
National University of Galway, Ireland
{brian.davis, fadi.badra, paul.buitelaar,tobias.wunner,
siegfried.handschuh}@deri.org

Abstract. An increasing number of enterprises are beginning to include ontologies into Text Analytics (TA) applications. This can be challenging for a TA group wishing to avail of such technologies due to the manual effort needed to map language resources within a TA system for a new domain. Ontology lexicalization offers a solution to this problem by seeking to automatically generate lexical resources in order to shrink the manual effort of this concept-to-text mapping process. However, conventional approaches are limited in that they often can only generate term mentions of proper noun, personal noun or fixed key phrases from concept labels in ontologies. Such approaches do not generalize to cope with more complex concept mentions such as nominal compounds or multi-word expressions. An alternative consideration is lemon - Lexicon Model for Ontologies which offers a more sophisticated solution to this problem. We describe a simple use case for exploiting lemon within a widely used open-source TA framework and demonstrate how lemon generated lexical resources are at least comparable in agreement to OntoRootGazeteer, a conventional ontology lexicalization approach.

Keywords: Ontology Lexicalization, NLP frameworks, Semantic Annotation

1 Introduction

An increasing number of enterprises are beginning to include semantic web ontologies into their Information Extraction and Text Analytics process regardless of whether this is to model the application domain or to model the internal data structures of text analytics system itself[1]. The Semantic Web/Linked Data community is also increasingly becoming aware of the need to encode linguistic knowledge concerning concepts directly into ontologies. In this paper we briefly describe lemon – Lexicon Model for Ontologies, which has been developed in the Monnet Project[2] in order to drive a standard for the sharing of lexical information across the semantic web. Furthermore we describe a simple experiment which uses an ontology based on

1 As demonstrated by the recent use of OntoText KIM for the BBC's 2010 World Cup.
2 http://www.monnet-project.eu/

food recipes. We generate a lemon lexicon model using existing services available from the Monnet website. Our goal is to demonstrate the ease of wrapping lemon API as a resource within widely used open-source framework – GATE[3][1]. Furthermore, we exploit lemon generated lexical resources for semantic annotation and provide a preliminary evaluation with promising results. The rest of this paper is structured as follows: Section 2 discusses the lemon model, the OntoRootGazeteer, which is an existing ontology lexicalization tool, distributed with GATE and key related work. Section 3 outlines our use case and implementation of a lemon resource in GATE, for the purpose of generating ontology aware lexical resources for semantic annotation. In Section 3, we compare the lemon approach with GATE's OntoRootGazeteer for observed agreement. Finally, Section 4 offers conclusions and future work.

2 Ontology Lexicalization – Tools and Related Work

Lemon – Lexicon Model for Ontologies

As mentioned earlier, lemon is a model sharing lexical information on the semantic web. Lemon is designed to be a:

- **Concise:** As small number of classes and definitions as needed.
- **Descriptive but not prescriptive**: it uses external sources for the majority of its definitions. A lemon based system can thus be extended in different ways for different tasks i.e. terminological variation, morpho-syntactic description, translation memory exchange.
- **Modular:** Lemon can be separated into a number of modules and it is not necessary to implement the entire lemon model to create a functional lexicon.
- **RDF-native:** Lemon is based on RDF for the purposes of interfacing and sharing across the semantic web. It also permits greater linking between different sections of the lexicon.

A simplest of a lemon entry is as follows:

```
@base <http://www.example.org/lexicon>

@prefix ontology: <http://www.example.org/ontology#>

@prefix lemon: <http://www.monnetproject.eu/lemon#>

:myLexicon a lemon:Lexicon ;

lemon:language "en" ;

lemon:entry :animal .
```

3 GATE - General Architecture for Text Engineering

```
:animal a lemon:LexicalEntry ;

lemon:form [ lemon:writtenRep "animal"@en ] ;

lemon:sense [ lemon:reference ontology:animal ] .
```

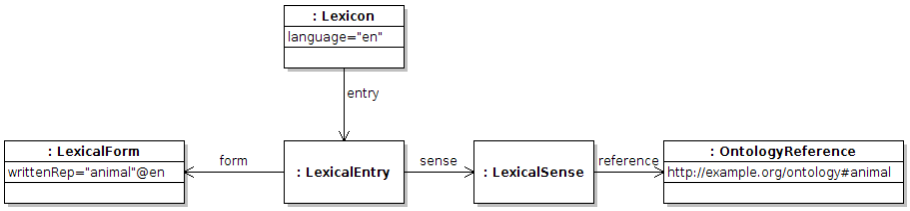

Fig. 1. Sample Lemon Entry visualized. (Extracted from lemon cookbook [4]).

Figure 1. defines the following entities:

1. `Lexicon` : This is the lexicon containing all elements in the lexicon. This approximately corresponds to a SKOS scheme.
2. `Lexical Entry` : This represents the given lexical entry.
3. `Lexical Sense` : Represents the relationship between the lexical entry and the ontology entity.
4. `Reference` : The reference to the resource that can be described by this lexical entry.
5. `Form` : A surface realization of a given lexical entry, typically a written representation

2.2 GATE OntoRootGazeteer

The goal of the GATE OntoRootGazeteer[5] is to produce ontology-based annotations i.e. annotations by pre-processing an ontology in order to extract human-understandable lexicalisations. The OntoRootGazeteer initially extracts all the names of ontology resources within a given ontology as well assigned property values for all ontology resources (e.g., label and data-type property values). Further processing involves replacing any name containing dash ("-") or underline ("_") character(s) with a blank space. In addition, built-in GATE lemmatizers and POS tagging resources are exploited in order to create the proper lemma for a given resource name. Finally an in-memory ontology aware gazetteer is created.

[4] http://www.monnet-project.eu/Monnet/resource/Monnet-Website/0000%20%20Library/0700%20-%20Downloads/lemon-cookbook.pdf. Accessed 15th August 2011

[5] http://gate.ac.uk/sale/tao/splitch13.html#x18-33900013.9

1.3 Additional Related Work

With respect to lemon, it is influenced strongly by Lexical Markup Framework- LMF [2] , which is part of the ISO TC37/SC4[6] working group on the management of Language Resources. LMF has its origins in language engineering standardization initiatives such as EAGLES[7] and ISLE[8] . With respect to in depth literature on lemon and its historical influences, we recommend [3] and [4]. The LIR (Linguistic Information Repository) model is similar in many respects, but focuses strongly on multilingualism [5]. Finally, there is OntoLing, which is a Protégé plug-in that allows for linguistic enrichment of ontologies[6].

3. Implementation and Experiment: Squeezing Lemon with GATE

3.1 Experimental Use Case

In our use case, we utilize an ontology of food recipes which contains a over four and half thousand classes of food ingredient. Currently it is unpopulated with instance data. We took the first one hundred concepts in the ontology for testing purposes. Our goal was not for scalability testing with respect to ontology storage but rather to test agreement between lemon generated lexical resources with those of the OntoRootGazeteer as well the ease of importing lemon as a resource into GATE.

3.2 LemonGazeteerGenerator PR

Using the online lemon generator[9], we uploaded our sample ontology to generate a a lexical model file in turtle[10] format. We wrote a small application using the lemon API, to iterate through all written representations for each given concept. For each unique concept in the lexicon mode, it creates a gazetteer list, which is a simple text file with lexical entries such as: *quail eggs and quail egg.* The application also writes an entry to a mapping definitions (See Figure 2) file which aligns ontology resources with gazetteer list entries. Finally, we wrapped the application as a GATE processing resource (PR) to promote language resource reuse (See Figure 3).

```
quail.lst:file:///home/bridav/Fadi/FoodOntologySmall.owl:Quail
avena.lst:file:///home/bridav/Fadi/FoodOntologySmall.owl:Avena
golden_raisin.lst:file:///home/bridav/Fadi/FoodOntologySmall.owl:Golden_raisin
foie_gras_entier.lst:file:///home/bridav/Fadi/FoodOntologySmall.owl:Foie_gras_e
```

6 http://www.tc37sc4.org/
7 http://www.ilc.cnr.it/EAGLES96/browse.html
8 http://www.mpi.nl/ISLE/
9 http://monnetproject.deri.ie/lemonsource
10 http://www.w3.org/TeamSubmission/turtle/

```
ntier
chanterelle.lst:file:///home/bridav/Fadi/FoodOntologySmall.owl:Chanterelle
```

Fig 2. *Mapping.def* : Contains alignments between gazetteer lists and ontology class URIs.

Once the following gazetteer list files and mapping file are created they can be exploited by the GATE OntoGazeteer resource, which is a hierarchical hash gazetteer for

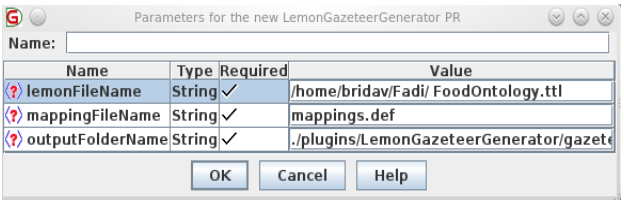

Fig 3. GATE LemonGazeteerGenerator PR, which takes a LemonModel turtle file as input and produces an ontology-aware gazetteer lists.

semantic annotation of concept mentions in text. However it differs from the OntoRootGazeteer in that it does not automatically lexicalize an ontology rather it follows traditional knowledge engineering approaches whereby the ontology must be manually aligned to lexical resources. In general it is used for small to medium sized ontologies where accuracy is critical, however for much larger ontologies it becomes unmanageable, hence the automatic approach using the OntoRootGazeteer. Note in our use case the OntoGazeteer is exploiting automatically generated lexical resources produced by our LemonGazeteerGenerator PR. Using existing GATE processing resources such as: a tokeniser, part of speech tagger and lemmatiser, we created a IE pipeline for semantic annotation. The same pipeline was reused with the OntoRootGazetteer to create annotations for agreement comparison. Recall the OntoRootGazeteer generates its own lexicalizations from the same food recipe ontology used by the LemonGenerator service.

3.3 Experimental Results

Using a small test corpus contain over 4650 lines of food recipes, we compared both the lemon generated OntoGazeteer and conventional OntoRootGazeteer.
As we do not at this time have a gold standard annotation set as a baseline, we only record observed agreement across both methods. Of the 798 annotations created by the OntoRootGazeteer, the LemonOntoGazeteer matched 74 % of annotations' spans,. Of those matches, 91% were in agreement with ontological concepts. Upon closer observation we noticed two issues:

1. The LemonGenerator web service appeared to have lexicalized only leaf node concepts in the food recipes ontology while the OntoRootGazeteer had traversed and lexicalized the entire graph. While this may seem a disadvantage. However, depending on the annotation task, it could be advantageous as an optional feature and thus benefit the user.

2. There were some unexpected errors in the LemonGazeteerGenerator PR in the form of some erroneous mappings. So for example, in addition to a mapping for `Fruit_Juice`, a mapping for a concept `Fruit` was also created. This may be a bug in either the PR or the LemonGenerator Service itself.
However despite these short comings, the lemon lexicalizations for the concepts mapped, were upon inspection correctly generated.

4 Conclusions and Future Work

In this paper, we have described initial experiments towards using lemon for ontology lexicalization. We demonstrated how easily lemon resources can be exploited by a well known TA framework. We compared a lemon based hierarchical gazetteer with the OntoRootGazeteer, a conventional ontology lexicalization tool available in GATE. We found that the results that results are at least comparable. The reader should note that we do not exploit the inherent multilingualism of lemon, nor its richer lexicalization features, which are not available to the OntoRootGazetteer. Future work will focus on exploiting the full power of the lemon model, improving the output of the lemon generator web-service as well as a more thorough evaluation.

Acknowledgments. : The work presented in this paper was supported (in part) by the European project MONNET No. (FP7/2007-2013) 248458 and (inpart) by the Lìon 2 project supported by Science Foundation Ireland under Grant No. SFI/08/CE/I1380

References

1. Cunningham, H., et al: Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science. (2011). ISBN 0956599311.
2. Francopoulo, G. George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C. : Lexical Markup Framework (LMF).In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC). Genoa, Italy, (2006).
3. McCrae, J, Spohr D, Cimiano P. : Linking Lexical Resources and Ontologies on the Semantic Web with lemon. Proceedings of the 8th Extended Semantic Web Conference (ESWC). Heraklion, Crete, (2011).
4. McCrae, J, Aguado-de-Cea G, Buitelaar P, Cimiano P, Declerck T, Gomez-Perez A, Gracia J, Hollink L, Montiel-Ponsoda E, Spohr D et al.: In Press. Interchanging lexical resources on the Semantic Web. Language Resources and Evaluation
5. Montiel-Ponsoda E, Aguado de Cea G, Gómez Pérez A, Peters W: Enriching Ontologies with Multilingual Information. Natural Language Engineering, (2010).
6. Pazienza,M., T., Stellat, A.: An Environment for Semi-automatic Annotation of Ontological Knowledge with Linguistic Content. In 3rd European Semantic Web Conference (ESWC 2006) Budva, Montenegro, (2006).