Denoting Data in the Grounded Annotation Framework

Marieke van Erp[1], Antske Fokkens[1], Piek Vossen[1], Sara Tonelli[2], Willem Robert van Hage[3], Luciano Serafini[2], Rachele Sprugnoli[2], and Jesper Hoeksema[1]

[1] VU University Amsterdam
{marieke.van.erp,antske.fokkens,piek.vossen,j.e.hoeksema}@vu.nl
[2] Fondazione Bruno Kessler {satonelli,serafini,sprugnoli}@fbk.eu
[3] SynerScope B.V. willem.van.hage@synerscope.com

Abstract. Semantic web applications are integrating data from more and more different types of sources about events. However, most data annotation frameworks do not translate well to semantic web. We describe the grounded annotation framework (GAF), a two-layered framework that aims to build a bridge between *mentions* of events in a data source such as a text document and their formal representation as *instances*. By choosing a two-layered approach, neither the mention layer, nor the semantic layer needs to compromise on what can be represented. We demonstrate the strengths of GAF in flexibility and reasoning through a use case on earthquakes in Southeast Asia.

1 Introduction

Semantic web applications are ingesting data from more and more different sources such as output from natural language processing applications, sensor data, videos or financial transactions. Each of these domains has their own data annotation practices which first need to be reconciled with semantic web standards. One issue with integrating information from different sources is that representation formats tend to look at their domain in isolation, making it difficult to integrate information that comes from other domains.

The Grounded Annotation Framework (GAF) [1] aims at addressing this problem by distinguishing instance *mentions* which can be domain specific from *instances* conform to domain independent semantic web standards. In this manner, we can integrate information for example extracted by NLP tools or from sensor data in a formal context which can be shared by different applications and over which we can perform reasoning. This paper addresses the advantages of using GAF from the point of view of users of Linked Data.

We will describe GAF in Section 2, present an example in Section 3 and conclude with pointers for future work in Section 4.

2 The Grounded Annotation Framework

The main property of GAF is that it distinguishes *instances* from *instance mentions*. A *mention* is the act of referring to an object where an *instance* is the object itself. The relation between instances and mentions is defined by *gaf:denotedBy*, which is the only new predicate GAF introduces. Different resources (or even the same resource) may refer to an instance in different ways and each of these references may have properties of its own. This is quite common in natural language, where authors tend to alternate terms to refer to the same object for stylistic reasons, but it can also play a role in other sources of information. If, for instance, a sensor displays a measured temperature, this displayed value has properties of its own that are clearly not properties of the value that was measured, such as the instrument that was used to measure it and its error rate. In the remainder of this contribution, we will illustrate GAF through the example of presenting instances in the Simple Event Model (SEM) [2] and mentions in the TERENCE Annotation Format (TAF) [3] which represents linguistic properties.

SEM is a model to express *who* did *what*, *where*, and *when*. It is not the only RDF model to describe events but as SEM is not tied to a any domain and is among the most flexible, we chose this model as the core of our semantic layer. It should be noted however that, in principle any RDF schema can be integrated into GAF. TAF is designed to annotate coreference relations between event mentions as well as participants, locations and temporal expressions, which covers the kind of information also represented in SEM. TAF has the additional advantage that it already distinguishes between *instances* and *instance mentions* for participants and locations. We use a slightly adapted variant of TAF that extends this distinction to events and temporal expressions as described in [1]. We chose TAF as it is based on the ISO-TimeML standard and fits our event use-case, however, any representation format can be used in GAF.

The *gaf:denotedBy* relations is used to link events represented in SEM to specific mentions represented in TAF. If a linguistic analysis identifies a syntactic relation between an event mention and the mention of a person, we can derive that this person is an Actor of the event in SEM according to the analysis of a specific text. Mentions thus play an important role in modelling **provenance** of information. To model provenance we use the PROV-O ontology [4] as it is compatible with our RDF representation and is recommended by W3C for provenance modelling. When we represent alternative views in SEM, these views are linked to the mentions they were derived from. This leads us to the original source and hence information in who expressed which view.

Creating GAF Annotations

GAF annotations can be created both by starting from the linguistic layer and the semantic layer. When starting from text for the mention layer, first TAF annotations are added to the text using the Celct Annotation Tool [5], which are then translated to SEM relations using a conversion script. Instances extracted from a particular source (for example a document) are grouped into named graphs, to which provenance information is added. We use manually defined rules for mapping TAF to SEM, but plan to use machine learning in the future.

When starting from the semantic layer, events and event properties are linked to textual mentions. We are currently working towards an annotation environment based on CROMER [6], which will allow the user to switch easily between the linguistic and semantic layers.

3 Examples

The example sentences shown in Figure 1 both contain information about the 2004 Indian Ocean Earthquake and Tsunami. The articles disagree on the cause of the earthquake; where Bloomberg ascribes it to moving tectonic plates, Veteran's Today sees a stealth attack submarine as the likely cause. Figure 2 shows that these two declarations can co-exist within the GAF representation of the earthquake. It is up to the application or user accessing the information to interpret the fact that there is a contradiction and for example select only particular sources for further processing. GAF provides the glue to connect non-semantic web data to semantic web representation formats. The *rdfs:isDefinedBy* relation at the top of Figure 2 shows how RDF predicates can be used to link GAF representations to external resources such as the Linked Open Data cloud.[1]

```
"Indonesia lies in a zone where the Indo-Australian, Eurasian, Philippine and Pacific plates
meet and occasionally shift, causing earthquakes and sometimes generating tsunamis. There
have been hundreds of earthquakes in Indonesia since a 9.1 temblor in 2004 caused a
tsunami that swept across the Indian Ocean, devastating coastal communities and leaving more
than 220,000 people dead in Indonesia, Sri Lanka, India, Thailand and other countries."
(Bloomberg, 2009-01-07 01:55 EST)

"...were most concerned about the cause, scope, and consequences of the December 26, 2004
Indian Ocean tsunamis because they were far bigger and more destructive than they had
anticipated. More important, it had no clear alibi that their most likely source of the
disaster, the Multi-Mission Platform of the new stealth attack submarine, the USS Jimmy
Carter, had not been the culprit."
(Veteran's Today, 2011-10-02)
```

Fig. 1. Sample sentences mentioning the December 2004 Indonesian earthquake

4 Conclusions and Future Work

We have presented GAF, a grounded annotation framework for integrating information from various sources. We have shown its flexibility in representing contradicting information from different textual sources.

We are currently developing an annotation tool that allows users to easily switch between linguistic and semantic annotation layers. After which we plan to develop tools supporting easy integration of other types of information, such as data from the Linked Open Data cloud, video metadata or sensor data.

[1] http://groundedannotationframework.org/ provides full examples and the GAF definition.

Acknowledgements
This research is supported by the European Unions 7th Framework Programme via the NewsReader Project (ICT-316404) and by the BiographyNet project, funded by the Netherlands eScience Center (http://esciencecenter.nl/).

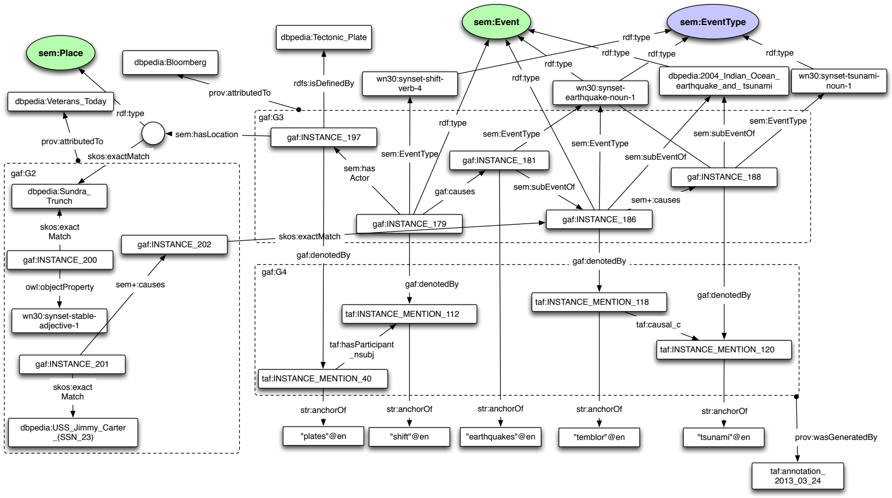

Fig. 2. GAF representation of Earthquake example

References

1. Fokkens, A., van Erp, M., Vossen, P., Tonelli, S., van Hage, W.R., Serafini, L., Sprugnoli, R., Hoeksema, J.: GAF: A grounded annotation framework for events. In: Proceedings of the first Workshop on Events: Definition, Dectection, Coreference and Representation, Atlanta, USA (2013)
2. Van Hage, W.R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G.: Design and use of the simple event model (SEM). Journal of Web Semantics (2011)
3. Moens, M.F., Kolomiyets, O., Pianta, E., Tonelli, S., Bethard, S.: D3.1: State-of-the-art and design of novel annotation languages and technologies: Updated version. Technical report, TERENCE project-ICT FP7 Programme-ICT-2010-25410 (2011)
4. Moreau, L., Missier, P., Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J., Miles, S., Myers, J., Sahoo, S., Tilmes, C.: PROV-DM: The PROV Data Model. Technical report, W3C (2012)
5. Bartalesi Lenzi, V., Moretti, G., Sprugnoli, R.: CAT: the CELCT Annotation Tool. In: Proceedings of LREC 2012. (2012)
6. Bentivogli, L., Girardi, C., Pianta, E.: Creating a Gold Standard for Person Cross-Document Coreference Resolution in Italian News. In: Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management. (2008)