Terminology-Based Patterns for Natural Language Definitions in Ontologies

Dagmar Gromann

Vienna University of Economics and Business, Austria
dgromann@wu.ac.at

Abstract. Natural language content in ontologies is crucial to any human interaction with them, but scarcely available. Terminology science centers on best practices in domain-specific natural languages. Hence, ontologies can benefit from the systematic approach of terminology to natural language definitions. This paper proposes an Annotation Ontology Design Pattern named "Natural Language Definition ODP" that provides natural language definitions for ontology classes. For this purpose, a (semi-)automated method for implementing this pattern combining ontology verbalization and information extraction is investigated herein and exemplified in the domain of finance.

Keywords: Annotation ODPs, Natural Language Definition, Terminology, Automatic Extraction of ODPs, Domain-Specific ODP Application

1 Introduction

A growing number of application scenarios for Semantic Web (SW) ontologies render reusable, high-quality solutions to their design increasingly important. For this purpose, Ontology Design Patterns (ODP) define a formal methodology for various aspects of ontological design, ranging from Logical to Presentation ODPs [1]. The latter seek to increase the usability and readability of ontologies from a user's perspective, which are vital to multi-lingual scenarios [2] and interactions with domain experts and users [3], and are divided into Annotation and Naming ODPs. Annotation ODPs provide best practices for homogeneous natural language (NL) expressions (`rdfs:label`) and definitions (`rdfs:comment`), while the latter focus on naming conventions [1]. A general paucity of an operational approach to NL definition authoring and its time-intensive nature led to scarce and frequently inconsistent NL definitions in ontologies. Thus, this paper investigates their (semi-)automated generation by means of the proposed Annotation ODP "Natural Language Definition ODP" based on established methods from terminology science.

Ever since its advent, terminology science has realized the need of providing a systematic approach to NL definitions. Concept-centered terminologies as defined by ISO 704:2009 and 1087:2000 consist of sets of terminology concepts in specialized domains. NL definitions are required to form these terminology

concepts and their interrelations. In contrast, ontology concepts are formally defined by means of logics. Combining the formal ontological definition with the terminological NL definition authoring method results in a multidimensional approach formalized as the proposed Annotation ODP. Thereby, the ISO 704 method of combining the denomination of the superordinate concept with (a) characteristic(s), delimiting the concept to be defined from its related concepts, can be (semi-) automated and is the foundation of the proposed pattern. Given domain-specific, axiomatized ontology elements with a minimum of NL coverage in labels or fragment identifiers, the superordinate concept's denomination can be identified and applied by using its subsumption hierarchy. For the non-trivial purpose of obtaining characteristics, three mutually complementing approaches are proposed: ontology verbalization, utilizing existing NL content, and information extraction. An example is provided by applying the pattern to the partially available NL content of Fadyart's Finance Ontology[1].

2 Natural Language Definition ODP

The objective of the NL Definition ODP is to define an ontology concept in natural language(s). Ontologies represent knowledge by formalizing vocabularies of terms as well as their interrelations and define their meaning formally. Terminologies mostly rely on NL characteristics to establish NL definitions for concepts and interrelate concepts designated by terms, appellations, or symbols. The two most important types of definitions as specified by ISO 704 are extensional definitions, listing the instances of a concept, and intensional definitions. The latter constitutes a combination of superordinate concept and manually identified delimiting characteristic(s) for concepts related generically.

The intensional approach offers the most explicit, consistent, and precise method to definition formation. It is intended to provide the minimum of information needed for human users to differentiate one terminology concept from another. To facilitate its automation, the basic textual description of ISO has been adopted and formalized for proposing the NL Definition ODP introduced in Definition 1 and illustrated in Example 1.

The pattern defines the NL definition of an `entry term`, which corresponds to the label of the ontology class. The singular form of the term is preferred, unless only available in plural, e.g. "liabilities". It utilizes the label or fragment identifier of the superordinate concept, which for the experiment herein is restricted to Noun Phrases (NP). Thereby, it obtains a context and implicitly inherits the characteristics of the superclass. The NP is connected to characteristics by utilizing a finite set of relative pronouns, verbs, and where applicable verbalized object properties. The same elements and a coordinating conjunction are needed to string together several characteristics.

Obtaining the characteristic(s) relies on a three-tiered mutually complementing approach of ontology verbalization, utilizing existing NL content, and infor-

[1] `http://fadyart.com/` version 3.04

mation extraction from structured Web resources. All three of them help specifying the relative pronoun and linking verb to be used for the concept to be defined.

Definition 1: NL Definition ODP

Entry Term [A/An] NP<superclass> [which/that/who/whose] [(can) be/include/belong to/classify as OR <objectProperty>] [(<characteristic(s)>)* and] <characteristic>

Example 1: NL Definition of Concept "Card" (Fadyart Finance Ontology)

Card [A] payment instrument [that] [has as card type] <a credit card or debit card> and [has as card data] <a start date, sequence number, holder name, expiry date, issuer name, card number, security code>

Ontology verbalization refers to the translation of ontology concepts, relations, and axioms to (controlled) natural languages, such as Attempto Controlled English [4]. In contrast to controlled natural language, the objective herein is to use verbalized ontology elements to identify the appropriate verb and relative pronoun linking the definition's characteristics. For this purpose, verbalization patterns have been identified, of which selected ones are provided below.

P1 - `ObjectUnionOf`: [a/an OR ObjectMinCardinalty] [(NP<class>,)* or] NP(class)
P2 - `ObjectMinCardinalty`: at least <number>
P3 - `ObjectSomeValuesFrom`: NP<class>(domain)[<ObjectSomeValuesFrom> that <ObjectProperty> [a/an] NP<class>(range(s))]
P4 - `ObjectProperty` with "has" is split into two parts: NP<class><(domain) <ObjectProperty:has> as <ObjectProperty:rest> [a/an] NP<class>(range)

Should the label of the object property already contain the concept label in the range, the concept label is not reiterated in the NL definition, e.g. "hasManager" pointing to "Manager". The above list is not exhaustive and requires NLP methods for its implementation, e.g. tokenization. The application of these patterns to characteristic formation will be exemplified in the next section.

In a next step, the existing `rdfs:comment` of the ontology class is linked to the NP and, where applicable, verbalized content by means of a coordinating conjunction and the identified relative pronoun, which, if not available in the comment, can be obtained from Wiktionary.

If no NL content is available, re-using existing structured Web resources, such as DBpedia, has been considered. The tentative information extraction process herein relies on string matching and an immediate subsumption to top DBpedia ontology concepts (e.g. Organization, Resource). Reducing NL definitions to DBpedia information might result in quality issues. For instance, circular definitions are frequent on DBpedia, i.e., a term is defined by itself or by a second term that refers back to the first term. For instance, "Debtor" is defined as "Debtor owes a debt to someone ...". Applying the proposed pattern ensures the proper context for the concept, i.e., superordinate concept, and DBpedia information provide useful additional details.

Due to its systematic nature, the described pattern enables a consistent formation of NL definitions, which strongly enhances the human readability of ontologies it is applied to. The proposed pattern is illustrated for the English language and requires minor adaptations for its realization in other NLs syntactically similar to English provided lexical resources are available.

3 Example Application

An OWL ontology serves as the input to the intended system design, here exemplified with Fadyart's Finance Ontology in English. By means of the OWL API the ontology can be parsed, the subclass relations and object properties identified, and an annotation property can be added. Starting from $\top$, the subsumption hierarchy is traversed to the first concept not directly subsumed by it. If its superordinate concept contains no label, its fragment identifier is tokenized (using e.g. the Stanford Core NLP) and represents the `NP<superclass>` of Definition 1. In Example 2, the class ClientPortfolio is the subclass of AccountsPayable.

To ensure the correct grammatical number and relative pronoun, the superclass term is queried in Wiktionary, e.g. Java-based Wiktionary Library[2]. Here, the query returns "plural only" and the relative pronoun "that"[3] for "Accounts Payable". Subsequently, tokenization and verbalization pattern P4 defined in the previous section are applied to the object property of Example 2. Its range consists of a union of three classes, which is verbalized using pattern P1. Finally, the existing comment is to be added to the already obtained definition. By means of Wiktionary "Clients" in the existing `rdfs:comment` is identified as countable noun, so its singular form can be combined with the obtained definition using a coordinating conjunction and the relative pronoun identified above. The derived definition can be added to the concept ClientPortfolio as `rdfs:comment`.

Example 2: Class ClientPortfolio in Manchester Syntax

Original Input in OWL

```
Ontology: <http://www.fadyart.com/Finance.owl> ...
ObjectProperty: hasClientPortfolioBeneficialOwnerOfIncome
  SubPropertyOf: hasAccountDomain
  Domain: ClientPortfolio
  Range: PartyHolder or PartyLegalRepresentative or
      PartyUsufructuary
Class: AccountsPayable
  Annotations: rdfs:label "Accounts payable"@en
  SubClassOf: ShortTermLiabilities
Class: ClientPortfolio
  Annotations: rdfs:label "Client accounts"@en,
  rdfs:comment "The clients of the financial institution for who's
      account the securities handling operations are performed."^^
      xsd:string
  SubClassOf: AccountsPayable
```

Resulting Definition

```
Client Account
Accounts payable that has as client
portfolio beneficial owner of income at
least one party holder, party legal
representative, or party usufructuary and
that is a client of the financial
institution for who's account the securities
handling operations are performed.
```

Several object properties and enormous unions render it necessary that domain experts decide which verbalization most adequately defines the concept. Additionally, at times comments are utilized for supplementary information rather than NL definitions of concepts, which is why for some cases the comments might not be re-used for the definition formation process.

4 Related Work

Glosses for ontology concepts reuse existing lexical resources, e.g. WordNet, to provide ontology engineers with various linguistic descriptions to choose from for a specific concept [5]. Approaches grounding existing ontologies in lexical and

[2] `http://code.google.com/p/jwktl/`

[3] "Money that is owed ..."

linguistic descriptions (e.g. *lemon*[4]) either re-use glosses for NL descriptions or derive meaning by pointing to the semantic object in the ontology. Ontology verbalization utilizes formalized knowledge in ontologies to derive NL descriptions. For instance, the SWAT project[5] facilitates the understandability of verbalized entailments by providing individual inference steps in the English language [6]. Instead of providing an external meta-model or ontology engineering support, the proposed pattern seeks to re-use existing resources and use verbalization patterns in order to provide NL definitions for existing domain-specific ontology classes. As a standard-based approach, it reflects established best practices and accepted semiotic theories.

5 Discussion and Future Work

This paper proposes a NL definition ODP on the basis of definition formation methods from terminology science. Subsequent to defining the pattern, a (semi-)automated design to obtaining NL definitions by means of ontology verbalization, utilization of existing NL comments, and information extraction has been exemplified in the financial domain. In terms of future work, the degree to which the pattern can be generalized to other domains will be tested. As regards, information extraction, a profound disambiguation process will be considered. Furthermore, its formalization for a submission to the ontology design pattern repository is planned.

References

1. Presutti, V., Blomqvist, E., Daga, E., Gangemi, A.: Pattern-Based Ontology Design. In Suárez-Figueroa, M.C., Gómez-Pérez, A., Motta, E., Gangemi, A., eds.: Ontology Engineering in a Networked World. Volume 12. Springer (2012) 35–64
2. Cimiano, P., Buitelaar, P., McCrae, J., Sintek, M.: LexInfo: A Declarative Model for the Lexicon-Ontology Interface. Web Semantics: Science, Services and Agents on the World Wide Web **9**(1) (2011) 29–51
3. Damljanovic, D., Agatonovic, M., Cunningham, H.: Natural Language Interfaces to Ontologies: Combining Syntactic Analysis and Ontology-Based Lookup through the User Interaction. In Sure, Y., Domingue, J., eds.: The Semantic Web: Research and Applications. Volume 19. Springer (2010) 106–120
4. Kaljurand, K., Kuhn, T.: A Multilingual Semantic Wiki Based on Attempto Controlled English and Grammatical Framework. In Corcho, P.C.O., Hollink, V.P.L., Rudolph, S., eds.: The Semantic Web: Semantics and Big Data. Volume 17. Springer (2013) 427–441
5. Jarrar, M.: Position paper: Towards the Notion of Gloss, and the Adoption of Linguistic Resources in Formal Ontology Engineering. In: Proceedings of the 15th international conference on World Wide Web, ACM (2006) 497–503
6. Nguyen, T.A.T., Power, R., Piwek, P., Williams, S.: Predicting the Understandability of OWL Inferences. In Corcho, P.C.O., Hollink, V.P.L., Rudolph, S., eds.: The Semantic Web: Semantics and Big Data. Volume 17. Springer (2013) 109–123

[4] `http://lemon-model.net`
[5] `http://swatproject.org`