Representing and Visualizing Text as Ontologies: A Case from the Patent Domain

Stamatia Dasiopoulou[1], Steffen Lohmann[2], Joan Codina[1], and Leo Wanner[1,3]

[1] Department of Information and Communication Technologies, Pompeu Fabra University, Barcelona, Spain
[2] Institute for Visualization and Interactive Systems, University of Stuttgart, Stuttgart, Germany
[3] Catalan Institute for Research and Advanced Studies, Barcelona, Spain

Abstract. This paper presents preliminary results on a framework for the representation and visualization of text as OWL ontologies under an open-domain paradigm, where no a priori schema for the facts to be extracted is available. The extracted ontology is visually represented as a specifically tailored node-link diagram. The applicability of the approach is demonstrated on a use case from the patent domain.

1 Introduction

Extracting ontologies from text can significantly facilitate knowledge integration and querying, through semantic alignment and mediation [6]. Only recently though, under the Linking Open Data (LOD) paradigm of publishing and linking structured information on the Web, has research shifted towards open-domain approaches, where no a priori schema for the facts to be extracted is available and the textual input is considered in its entirety [1,20].

Within such context, two main challenges emerge: 1) to ensure the translation of textual input into well-formed ontologies that facilitate knowledge integration and querying in a schema-agnostic fashion; and 2) to provide the means for comprehensive visualizations that foster the understanding of the extracted knowledge, particularly at the factual level, in an intuitive manner that appeals adequately to users of diverse backgrounds and with varying levels of expertise.

These challenges are sharply manifested in the patent domain. The highly specialized and cross-domain terminology used in patent documents makes it very difficult, if not impractical, to rely on the availability of predefined schemata for the extraction of knowledge relevant to the task at hand. Moreover, the inherent complexity of patent documents render effective visualizations key tools for assisting experts in quickly grasping the main elements and their interactions.

In this paper, we present preliminary results on a framework for the representation and visualization of text as an OWL ontology under an open-domain paradigm, and illustrate its application with a use case from the patent domain. Abstracting from the specifics of the various semantic parsing methodologies, we describe an entity-relation-centric model for OWL-based text representation together with a graphical notation for its visualization as node-link diagram.

2 Related Work

In accordance with the twofold goal of the proposed framework, related approaches to ontology extraction and visualization are discussed in the following.

2.1 Extracting Ontologies from Text

Although ontology learning and population from text have been the subject of arduous research [4,21], investigations into the conceptualization of text in its entirety have commenced only recently with LODifier [1] and FRED [20]. Both use Boxer [7] to extract Discourse Representation Structures (DRSs), namely *discourse referents* (entities) and *conditions* (unary and binary relations), and respective rules to translate them into ontological representations. LODifier keeps modeling commitments minimal, by introducing a blank node for each discourse referent and by using reification to capture embedded DRSs. FRED [20] implements a more earnest mapping of DRSs to OWL constructs, utilizing frame semantics [2], links to the DOLCE+DnS foundantional ontology and heuristic rules that aim to maximize conformance to Semantic Web best practices.

Both result in representations that explicitly cater for *n-ary* relations, which represent a critical share of relations for effectively capturing the richness of textual contents. However, LODifier compromises ontology design with choices such as blank nodes, whereas FRED ensures high compliance with best practices, but the presented translations and heuristic rules are specifically tailored to DRSs. Instead, our goal is to provide a model for the generation of OWL representations from text that avoids commitments to specifics of the predicate-argument structures.

2.2 Visualizing Fact-based Ontologies

Many approaches to graphically represent ontologies have been proposed in the last couple of years [8,14]. However, they are not tailored to the visualization of ontologies that are extracted from text, and have limitations in this regard. While some approaches (e.g., OWLViz [13] and KC-Viz [18]) merely visualize the class hierarchy of ontologies, others (e.g., OntoGraf [10] and FlexViz [11]) are able to represent different types of properties. All these attempts are related to the visualizations generated by FRED in that they focus on terminological knowledge (aka TBox) and not on assertional knowledge (aka ABox), which we aim to visualize in our work. The same holds for ontology visualizations that provide more elaborated notations (e.g., Graffoo [9] and VOWL [15]), i.e., they also mainly address the ontology schema and are therefore less appropriate for the representation of fact-based ontologies extracted from text.

This is different in visualizations of RDF and Linked Data that are typically more oriented towards the ABox. Examples include RDF Gravity [12], Welkin [17], and LodLive [5]. Such visualizations depict the triple structure of RDF but they are usually not capable to represent n-ary relations. In addition, they use plain node-link diagrams with only little variation in the visual elements.

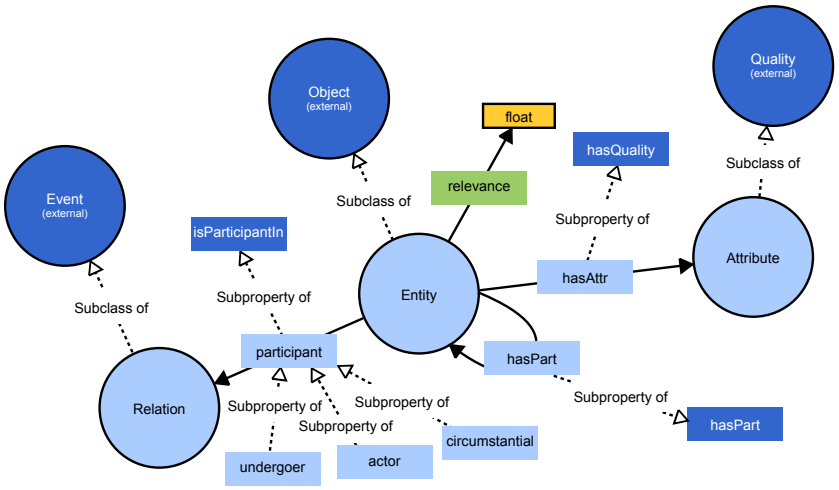

Fig. 1. Classes and properties of the core vocabulary (visualized with VOWL).

3 Ontological Text Representation

Aiming to abstract from predicate-argument specifics while assuring maximal interoperability within the Semantic Web and LOD context, we developed a minimal reference model for generating ontological text representations at a factual level.[4] Hence, our goal is to provide core classes and properties for capturing the ways in which the extracted entities are interrelated, and that can be applied across domains, serving as anchors for attaching application-tailored class and property hierarchies.

A key design decision has been to model the extracted relations as classes rather than properties. This is motivated by the saliency of n-ary relations in textual resources, and the incurred loss of semantics when, instead of preserving the n-ary dependencies, they are broken down into binary relations [19]. Furthermore, direct mappings to well-established foundational ontologies, such as DOLCE+DnS Ultralite[5] and SUMO[6], are promoted to enhance the interoperability and compliance with ontology design practices.

In accordance with the aforementioned principles, the model comprises the following core classes: `Entity` subsumes the set of physical objects, processes, and substances; `Relation` captures n-ary interrelations between entities; `Attribute` encompasses characteristic aspects of an entity that cannot exist without it. Alongside, a minimal set of upper-level object properties connect individuals of the three classes: `participant` allows to link entities to the relations in which

[4] Modalities, such as belief, causality, and entailment, are not considered as they can be covered through specialized ontologies and knowledge patterns.
[5] `http://www.ontologydesignpatterns.org/ont/dul/DUL.owl`
[6] `http://www.adampease.org/OP/SUMO.owl`

they participate; `actor` and `undergoer` specialize `participant` in order to discriminate between direct participants ("who?", "what?", etc.) and complementary ones (e.g., the $[\text{pump}]_{actor}$ pumps $[\text{water}]_{undergoer}$), while `circumstantial` is a specialization used as a catch-all property for other types of participation; `hasAttr` is used to associate entities to their attributes; `hasPart` is used to capture mereological relations between entities; lastly, the datatype property `relevance` allows to capture the relevance of the extracted entities to the matter being considered. Figure 1 visualizes the core vocabulary using VOWL [15]. The vocabulary is aligned with classes and properties from the DOLCE+DnS Ultralite ontology, which have a white font on a dark background in Figure 1.

The extracted predicate-argument structures can then be translated into OWL representations, according to the following rules:

- For each extracted entity, attribute, or relation, a named individual is generated; for co-referential entities, i.e., entities referring to the same real-world object, a single individual is introduced.
- For each added named individual, respective `rdf:type` statements are added based on the extracted vocabulary of entities, attributes, and relations.
- Respective `rdfs:subClassOf` axioms are added for each introduced entity, relation, and attribute class.
- Instigative and passive participation links between entities and relations are translated into respective `actor` and `undergoer` property assertions; likewise for circumstantial participation, where additionally the prepositions lexicalizing the participation are defined as subproperties. For example, given the excerpt "...connected along...", `along` is added as a subproperty of `circumstantial`.
- Links between entities and attributes as well as entities and their parts are captured as `hasAttr` and `hasPart` property assertions, respectively.

The result is an OWL ontology consisting primarily of assertional knowledge, i.e., class and object property assertions, and to a lesser extent of terminological knowledge, as it could be derived from links to LOD resources, such as DBpedia and WordNet. Further specializations and schema enrichments, according to the given application needs, can be acquired through ontology learning.

4 Visualization of the Extracted Ontology

Our visual notation for the graphical representation of the extracted ontology is inspired by VOWL [15], which provides user-oriented visualizations for OWL ontologies. VOWL has, for instance, been used to create the visualization of Figure 1. However, whereas VOWL focuses on the visualization of the ontology schema, we are interested in the visualization of facts extracted from text. Therefore, we could not simply reuse VOWL but developed a related ABox visualization that combines the strengths of VOWL with the peculiarities of visualizing fact-based ontologies extracted from text.

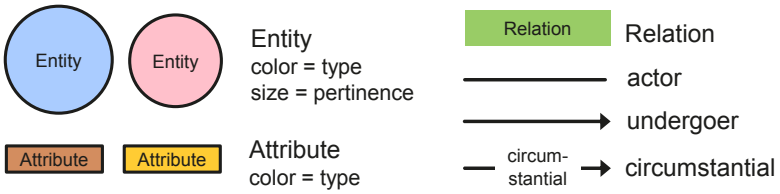

Fig. 2. Notation for the graphical representation of the extracted ontology.

Figure 2 summarizes the current visual notation. We adopted the basic visual elements of VOWL, consisting of circles which represent the extracted entities and rectangles representing the relations. The colors of the circles and attributes can be varied depending on their type. In contrast to VOWL, relations can be n-ary, which requires that they are rendered as nodes. This is in line with our design decision to model relations as classes rather than properties in the extracted OWL ontologies. Furthermore, we introduced a labeled link element to depict prepositions that qualify circumstantial participations.

We also adopted the idea of scaling the size of the circles, which, in VOWL, reflects the number of individuals that are members of a class. In our case, the circle size indicates the relevance values computed for the terms: Entities with a higher relevance value are shown in a larger size in the visualization. This helps to easily spot those entities that are most relevant to the matter being considered.

Finally, we decided to attach the attributes directly to the entity nodes instead of adding another link, as for the datatype properties in VOWL, in order to emphasize their strong connection and visually indicate that attributes cannot exist without the corresponding entities.

5 Use Case from the Patent Domain

Patent documents are highly idiosyncratic, verbose texts that describe elaborate inventions and make heavy use of specialized terminology. These characteristics, in combination with the continuously growing rate at which patents are filed worldwide, incur extensive labor and time costs for carrying out typical patent portfolio analysis tasks. In this context, structured representations that can assist experts in identifying and contrasting patents relevant to the task in question, by rendering semantics explicit, and visualizations that effectively summarize the key elements of an invention and foster understanding, can entail immediate competitive advantages.

In the investigated use case, we address constructive patents, i.e., patents that describe the constituent parts of machine inventions and the ways in which they interact. In this context, it is important to specialize the described entities into components (e.g., *coil*, *battery*), substances, processes, and other entities (e.g., *temperature*); likewise, for spatial and quantity attributes, such as *inner* charger

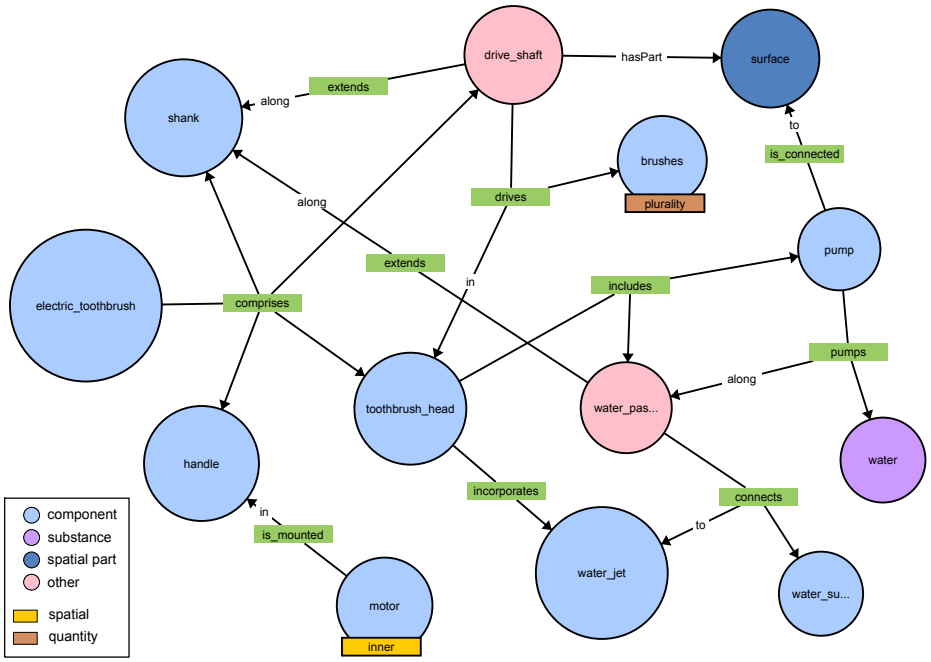

Fig. 3. Visualization of an ontology extracted from the claim text of a patent.

and *plurality* of brushes, as well as spatial parts (e.g., *surface*, *bottom*). To this end, the upper-level model definitions have been extended accordingly through the introduction of respective subclasses to the classes `Entity` and `Attribute`.

Using the *mate tools* [3], predicate-argument structures are extracted and subsequently their relevance is computed following a methodology similar to one used for identifying relevant sentences in extractive summarization tasks [16]. Then, OWL representations in compliance with the extended core model are generated, based on the transformation rules described in Section 3.

Figure 3 shows the visualization resulting from the below patent claim, where the extracted entity, relation, and attribute individuals are outlined in respective fonts. The initial layout of the diagram has been generated with a force-directed algorithm and has then been manually adapted to increase its readability.

An ***electric toothbrush*** *with a* ***water jet****, the* ***toothbrush*** COMPRISING *a* ***handle****, a* ***shank****, and a* ***toothbrush head*** *that* INCORPORATES *the* ***water jet****, in which an inner* ***motor*** IS MOUNTED *in the* ***handle*** *and the* ***toothbrush*** INCLUDES *a* ***reciprocating drive shaft*** EXTENDING *along the* ***shank*** *to* DRIVE *a plurality of* ***brushes*** *in the* ***brush head****,* INCLUDING *a* ***water passage*** EXTENDING *along the* ***shank*** *to* CONNECT *a* ***water supply*** *to the* ***water jet****, and a* ***pump*** *in the* ***shank*** *for* PUMPING ***water*** *along the* ***passage*** *that* IS *mechanically* CONNECTED *directly to the surface of the* ***drive shaft****.*

In the given example, there are four types of extracted entities (components, substances, spatial parts, and other) and two types of attributes (spatial and quantity), as indicated by the different colors assigned to the entity and attribute

nodes. As mentioned before, coreferential entities are captured by a single individual, upon which the respective participation links are projected. For example, the mentions of passage in "...a `water passage` EXTENDING along the `shank`..." and "...PUMPING `water` along the `passage`..." refer to the same passage entity; accordingly, there is a single "water passage" node to which the participation links in the EXTENDING and PUMPING relations have been projected.

All in all, the visualization provides an adequate representation of the patent claim that could be used to support analysts in understanding the elements and interrelations of the described invention.

6 Conclusions and Future Work

In this paper, we have presented an upper-level model for extracting ontological text representations under an open-domain paradigm that allows abstracting from the specifics of predicate-argument structures, and a visual notation for its graphical representation that focuses on the visualization of facts rather than the ontological schema. The applicability of the proposed representation and visualization framework has been demonstrated through a use case from the patent domain.

Future work includes further validation and fine-tuning of the representation model, through extensive evaluation in cooperation with experts from the patent domain, as well as in an application-wise manner, where it will be used as the basis for assessing semantic similarity between patents. Furthermore, future research will have to address enhanced visualization paradigms that are more tailored to the patent domain.

General challenges with regard to the notation are improved scalability and readability of the visualization. A scalable visualization must be capable to represent larger ontologies extracted from several paragraphs of a text. In the patent use case, the individual claims could, for instance, form different subgraphs that are connected with each other according to specified dependencies.

Generalizing from the patent domain, the presented representation and visualization framework may serve as a valuable starting point for related cases of ontology extraction and visualization. The open-domain character of the ontology extraction and representation approach enables its wide application, along with the visual notation that combines the clarity of VOWL with an ABox-oriented view and capabilities to explicitly represent n-ary relations.

Acknowledgments

This work has been supported by the EU FP7-SME-606163 project iPatDoc.

References

1. Augenstein, I., Padó, S., Rudolph, S.: LODifier: Generating linked data from unstructured text. In: 9th Extended Semantic Web Conference (ESWC '12). pp. 210–224. Springer (2012)

2. Baker, C., Fillmore, C., Lowe, J.: The Berkeley FrameNet project. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th Int. Conference on Computational Linguistics (COLING-ACL '98). pp. 86–90. ACL (1998)
3. Bohnet, B., Nivre, J., Boguslavsky, I., Farkas, R., Ginter, F., Hajic, J.: Joint morphological and syntactic analysis for richly inflected languages. Transactions of the Association for Computational Linguistics 1, 415–428 (2013)
4. Buitelaar, P., Cimiano, P. (eds.): Ontology Learning and Population: Bridging the Gap Between Text and Knowledge. IOS Press (2008)
5. Camarda, D.V., Mazzini, S., Antonuccio, A.: LodLive, exploring the web of data. In: 8th International Conference on Semantic Systems (I-SEMANTICS '12). pp. 197–200. ACM (2012)
6. Cimiano, P.: Ontology learning and population from text - algorithms, evaluation and applications. Springer (2006)
7. Curran, J., Clark, S., Bos, J.: Linguistically motivated large-scale NLP with c&c and boxer. In: 45th Annual Meeting of the Association for Computational Linguistics (ACL '07). ACL (2007)
8. Dudáš, M., Zamazal, O., Svátek, V.: Roadmapping and navigating in the ontology visualization landscape. In: 19th International Conference on Knowledge Engineering and Knowledge Management (EKAW '14). pp. 137–152. Springer (2014)
9. Falco, R., Gangemi, A., Peroni, S., Shotton, D., Vitali, F.: Modelling OWL ontologies with graffoo. In: ESWC 2014 Satellite Events. pp. 320–325. Springer (2014)
10. Falconer, S.: OntoGraf. http://protegewiki.stanford.edu/wiki/OntoGraf (2010)
11. Falconer, S., Callendar, C., Storey, M.A.: A visualization service for the semantic web. In: 17th International Conference on Knowledge Engineering and Knowledge Management (EKAW '10). pp. 554–564. Springer (2010)
12. Goyal, S., Westenthaler, R.: RDF Gravity. http://semweb.salzburgresearch.at/apps/rdf-gravity/ (2004)
13. Horridge, M.: OWLViz. http://protegewiki.stanford.edu/wiki/OWLViz (2010)
14. Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., Giannopoulou, E.: Ontology visualization methods – a survey. ACM Computer Surveys 39(4) (2007)
15. Lohmann, S., Negru, S., Haag, F., Ertl, T.: VOWL 2: User-oriented visualization of ontologies. In: 19th International Conference on Knowledge Engineering and Knowledge Management (EKAW '14). pp. 266–281. Springer (2014)
16. Mani, I.: Automatic summarization. John Benjamins Publishing (2001)
17. Mazzocchi, S., Ciccarese, P.: Welkin. http://simile.mit.edu/welkin/
18. Motta, E., Mulholland, P., Peroni, S., d'Aquin, M., Gomez-Perez, J.M., Mendez, V., Zablith, F.: A novel approach to visualizing and navigating ontologies. In: 10th International Semantic Web Conference (ISWC '11), Part I. pp. 470–486. Springer (2011)
19. Noy, N., Rector, A., Hayes, P., Welty, C.: Defining n-ary relations on the semantic web. http://www.w3.org/TR/swbp-n-aryRelations/ (2006)
20. Presutti, V., Draicchio, F., Gangemi, A.: Knowledge extraction based on discourse representation theory and linguistic frames. In: 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW '12). pp. 114–129. Springer (2012)
21. Wong, W., Liu, W., Bennamoun, M.: Ontology learning from text: A look back and into the future. ACM Computer Surveys 44(4) (2012)