Journal of Universal Computer Science, vol. 24, no. 11 (2018), 1651-1676
submitted: 28/11/17, accepted: 7/9/18, appeared: 28/11/18 © J.UCS

Human Language Technologies: Key Issues for Representing Knowledge from Textual Information

Yoan Gutiérrez
(University of Alicante, Spain
ygutierrez@dlsi.ua.es)

Elena Lloret
(University of Alicante, Spain
elloret@dlsi.ua.es)

José M. Gómez
(University of Alicante, Spain
jmgomez@ua.es)

Abstract: Ontologies are appropriate structures for capturing and representing the knowledge about a domain or task. However, the design and further population of them are both difficult tasks, normally addressed in a manual or in a semi-automatic manner. The goal of this article is to define and extend a task-oriented ontology schema that semantically represents the information contained in texts. This information can be extracted using Human Language Technologies, and throughout this work, the whole process to design such ontology schema is described. Then, we also describe an algorithm to automatically populate ontologies based our Human Language Technology oriented schema, avoiding the unnecessary duplication of instances, and having as a result the required information in a more compact and useful format ready to exploit. Tangible results are provided, such as permanent online access points to the ontology schema, an example bucket (i.e. ontology instance repository) based on a real scenario, and a documentation Web page.
Key Words: Ontology Development, Ontology population, Human Language Technologies, Semantic Package, Knowledge Engineering
Category: H.2.3, H.3.3, M.0

1 Introduction

The vast amount of available information is impossible to manage without the help of automatic tools and applications. In this respect, search engines or question answering systems have become essential in our everyday lives. The creation of these tools is possible thanks to the research and development conducted into Human Language Technologies (HLT) [Varile and Zampolli, 1997]. However, the information, and in particular, textual information, continues increasing at an exponential rate with the particular feature that related information about the same topic/issue/entity is normally dispersed throughout different documents and not connected. Moreover, although HLT applications are very useful to process and extract information easier and faster, their output normally solves a

specific task (e.g., disambiguate [Gutiérrez et al., 2010] or summarize [Lloret and Palomar, 2012]). This results in a lack of interconnection between the different outputs, which minimizes the potentials of combining all the outputs together, and therefore the global understanding of a text.

Linguistic analyzers or HLT frameworks allow to run different types of analysis over a text; but they do not keep or connect the analysis carried out after processing several texts. In contrast, ontologies can be used to keep the information interconnected. This is the case of DBpedia [Lehmann et al., 2012] that connects all the entities in the Wikipedia with respect to their relationships by using ontologies and RDF triples. In addition, they can model and represent the semantics of a broad range of domains, in order to further inferring and/or reasoning knowledge about that domain (e.g., tourism [Chaves et al., 2012], financial [Krieger et al., 2012] and others), and purposes (e.g., interoperability [Suca and da Silva, 2013], classification [Costa et al., 2013] and others).

Although ontologies are normally used for the conceptual modeling of a specific domain, they have a greater potential to be used in wider contexts that are still unexplored. For instance, they can be used to capture the semantics of a document written in natural language regardless the domain the document belongs to, thus being able to interlink heterogeneous documents spread by the Internet, and infer new meaning from them. This is not an easy task and given this context, the main goal of this research article is twofold: (i) to define and extend an ontology schema that can be used for representing textual information and derive new and implicit knowledge; and (ii) to validate and show the usefulness of the ontology within a case of study by means of 30 competence questions, aided by an example ontology repository built for that purpose. In addition, a detailed algorithm for automatically processing documents and populating the ontology repository is provided. To the best of our knowledge, no previous research work has been proposed to define such a general-purpose ontology to advance the state of the art in this field, regardless the domain or the purpose for which an ontology needs to be defined. Only in [Lloret et al., 2015], this idea was stated by proposing a preliminary small ontology schema. However, it had several limitations concerning the definition of concepts and relations and the management of duplicated information.

As a result, a final stable version reusing and linking information from/to other existing ontologies (e.g., DBpedia, and others) has been made available. Therefore, reusing HLT processes on top of an appropriate ontology contributes to the better understanding and representing a text, also allowing the development of more flexible semantic resources.

2 Related work

For building ontologies, several methodologies are available, such as BSDM [IBM, 1990], which provides the guidelines developed by IBM for modelling enterprises as a preliminary step for developing IT systems; the one proposed by Uschold and King in 1995 called METHONTOLOGY [Uschold and King, 1995], which is one of the most comprehensive methodologies available for building ontologies; KADS [Tansley and Hayball, 1993], a structured way of developing knowledge-based systems (expert systems); IDEF5 [KBSI, 1994], a software engineering method to develop and maintain usable, accurate, domain ontologies; and Tom Grubers principles for ontology design [Gruber, 1995], an engineering perspective on the ontology development.

There are different studies that deal with ontologies for Human Language Technologies semantic representation. For example, NLP Interchange Format (NIF) [Hellmann et al., 2013]. NIF is based on a Linked Data enabled URI scheme for identifying elements in (hyper-) texts and an ontology for describing common natural language processing terms and concepts. NIF enables the creation of heterogeneous, distributed and loosely coupled HLT applications, which use the Web as an integration platform. Due to the great relationship of NIF with our proposal, we have reused and integrated some of its terms. Similar woks can be found in the lexicon model for ontologies (lemon[1]) as a main outcome of the work conducted by the Ontology Lexicon community group (Ontolex). To some extent, this ontology addresses the lack of support for enriching ontologies with linguistic information, and more specifically, with information concerning how ontology entities can be realized in natural language. Instead of that, our work starts by collecting all possible conceptualizations from documents focused on final HLT tasks, lemon research does not provide at this level interesting RDF concepts useful to be reused by us.

In [Rospocher et al., 2016], a bottom-up approach to automatically build knowledge graphs from news articles using HLT tools and resources is proposed, representing main events happening in documents together with all their associated information. They do not consider, for example, the subjectivity of a statement, which is relevant in our approach. In contrast, ours follow a top-down approach, where a domain- and genre-independent ontology is first designed to capture the meaning behind a document, and then, its concepts and relations are instanced based on the information detected and extracted through HLT tools.

For populating ontologies, the standard methodology proposed consists of three tasks: i) candidate instances identification; ii) classifier construction, and iii) instance classification. The main challenge is to achieve a methodology that

[1] `http://cimiano.github.io/ontolex/specification.html`, last access Jun 2018

is domain-independent to reduce the time and cost of ontology instantiation.

In [Celjuska and Vargas-vera, 2004], an approach called Ontosophie for semi-automatic population of ontologies with instances from unstructured text is proposed. Supervised learning techniques are employed to learn extraction rules from annotated text and then apply those rules on newly articles for ontology population. The approach is based on three components: a natural language processing component, a dictionary induction tool, and an information extraction component. User interaction is required to take the final decision about the extracted instances. Therefore, in the end, manual intervention from users is needed to validate the suggested instances.

Another approach that addresses the automatic population of an ontology with named entities can be found in [Shen et al., 2012]. In this case, the instances of the ontology will be limited to only named entities. Given a named entity mention detected from the unstructured text, if the mapping entity of the mention is not contained in the ontology, the right category node to which the entity mention should be attached needs to be found. Otherwise, if the entity already exists in the ontology, the aim of this task would be to link this detected mention with its corresponding real world entity in the ontology (known as the entity linking task). Its main limitation is the lack to address other key elements that may be also present in texts, such as concepts, subjective sentences, topics, or domains.

Existing approaches that reuse natural language processing tools to extract information from text documents to populate ontologies can be found in [Draicchio et al., 2013, Faria et al., 2014, Corcoglioniti et al., 2016, Basile et al., 2016].

All previous approaches are limited in the type of elements that can be considered as instances, since they only rely on very specific processes, such as morpho-lexical analysis, named entities extraction or co-reference resolution. Moreover, the ontologies behind these processes are not clearly defined or validated neither how and to what extent the relationships between the extracted candidate instances are obtained. In this manner, the discovery and annotation of relationships between instances in the ontology is a very important stage to allow knowledge inferring from the ontology in further processes.

The novelty of our research lays on the fact that the text is seen as a whole object with the identification of implicit and explicit information, as candidate instances. These can range from named entities to concepts, but also involving summaries, sentences, sentiments or domain information. Avoiding duplicating unnecessary instances is one of the most important issues to be taken into consideration when automatically populating ontologies. We therefore create an ontology schema that captures and represents this, as well as describing an algorithm proposal for automatic populating it in a feasible and reliable manner.

3 Motivation and usefulness of an HLT-oriented ontology

3.1 Why are ontologies needed for representing texts?

Texts can be characterized by specific features that distinguish them from a set of disconnected sequence of sentences. The most evident elements a text contains are *words* that are connected to form *sentences*. But, when a text is analyzed in-depth, other types of elements can be found, discovering valuable information that is not explicitly stated in the text. This implicit information may comprise *dates* (e.g., "2nd March 2016", "yesterday", "the next Sunday"), *named entities* (e.g. "Spain", "Barack Obama", "Starbucks"), or even *word senses* (e.g., "bank", with the meaning of financial institution). Also, implicit meaning from sentences can be also obtained. This includes the domain a word/sentence belongs to (e.g., "sports"), the polarity/sentiment of a word/sentence (e.g., positive for the sentence "I like very much this place"), the type of named entities (e.g. entities of person, organisation, place, etc.), or the gist of the document (e.g. its summary), which imply a deeper understanding and reasoning process. When a human reads the text, the previous types of elements and semantic information may be easily detected and understood, but when the task of text processing is carried out automatically, it becomes much more difficult. In this sense, HLT tools can be used for identifying and extracting different information implicitly or explicitly stated in a document.

In this context, if we want to represent and integrate all this information together, the use of ontologies is very appropriate, where each type of information would be represented as a concept in the ontology [Lloret et al., 2015], and then the relations between them would be provided, as it is explained in the next section. Once the ontology is ready, the output of HLT tools would constitute the instances to populate the ontology.

3.2 How to use and represent textual information?

In order to reuse the information provided by the HLT analysis on documents, it is necessary to design and develop a task-oriented ontology schema. In this manner, it would be possible to capture the semantics of a document, taking into account the linguistic phenomena a text can include, and automatically populating it using the output of different HLT tools. Moreover, in the design process it is also necessary to consider the specific type of users that will later consume the ontology repositories. In our case, these users could be HLT **experts** or **data analysts**. The former would be interested in making the use of the ontology more extensible by extracting multiple lexical and semantic data included in the documents from which the ontology will be populated. Data analysts would exploit the whole ontology for extracting many combinations of semantic queries in order to generate reports based on concurrent evidences.

The ontology schema developed should be able to provide enough conceptual representations to capture the semantics of documents through a set of key aspects in texts, such as the temporal dimension (i.e. date mentions), presence of named entities, detection of opinionated information (i.e. positive or negative judgments), or conceptual classifications (i.e. document categories like sports, medicine, etc.). In addition, it should provide a lexical dimension, where we can represent the sentences of each document, and a possible summary derived from it. All these issues are crucial for setting up our own interpretation of possible scenarios (a meta-level specification) and vocabulary to consider.

Due to the fact that our main purpose is to provide meta-analysis specifications of documents, which could be reusable by a large community in a standard form, we planned to establish basic HLT terminology outlined by experts in this research field and then formalize them in a document-oriented ontology schema. This way, we will be able to automatically generate instances as persistence stage of document processing.

4 Ontology engineering

In this research we opted for METHONTOLOGY [Uschold and King, 1995] since it is the most suitable for developing task-oriented ontologies. The framework enables the construction of ontologies from the knowledge level (i.e., the conceptual level) to the implementation level, proposing a development life cycle, techniques, outcomes and evaluation principles for implementing ontologies.

Our development life cycle was carried out by means of the following tools: Protégé[2], Protégé visual plugins (OWL Viz[3], Ontograf[4], VOWL[5], among others) and query language tools [Sirin and Parsia, 2007] (e.g. SPARQL, DL Query).

It is important to note that the enhanced and final ontology schema of this research work, Semantic Package version 1.1, is an extension of a previous work described in [Lloret et al., 2015] to address the limitations found after the analysis conducted. These limitations were: some terms and relationships required re-factorization; related ontologies were not reused; lack of ontology metrics; lack of online accessibility and documentation; and lack of online example where SPARQL queries were tested. Therefore, with the goal to advance forward and create solid bases on this research field, this research work addresses these drawbacks in order to improve its ontology design to better capture and represent the information in texts.

[2] `http://protege.stanford.edu/`, last access Jun 2018
[3] `https://protegewiki.stanford.edu/wiki/OWLViz`, last access Jun 2018
[4] `https://protegewiki.stanford.edu/wiki/OntoGraf`, last access Jun 2018
[5] `https://protegewiki.stanford.edu/wiki/VOWL`, last access Jun 2018

4.1 Reuse and integration

In terms of reusing other shared ontologies, Semantic Package version 1.0 was redesigned to 1.1 mostly according to the semantic vocabulary defined in DCMI[6] (Dublin Core Metadata Initiative) and DBPedia[7], to represent metadata and linked data of, for example, different simple and generic resource descriptions but conserving the class hierarchy defined in [Lloret et al., 2015]. Furthermore, in the version 1.1 we have made the following improvements: re-factorization of terms and relationships; re-usability of shared ontologies; detailed evaluation by considering different ontology metrics; an algorithm for automatic populating the described ontology schema, testing and simulation of a real scenario, accessibility to a permanent public documentation[8]; accessibility to the permanent schema of the Ontology[9]; and accessibility to a permanent example repository with instances of the scenario described in this work[10].

Most of the metadata in our ontology schema was aligned to external sources of the Semantic Web to achieve the level 1 of interoperability that propose these shared sources Table 1). Other terms came from NIF[11] [Hellmann et al., 2013] and WordNet RDF[12]. As it can be seen in Table 1, some initial concepts were aligned with concepts of shared ontologies which entailed the variation of the initial names. The affected terms can be found in Table 2. In our ontology, we have chosen verbs to describe relationships, while in DBpedia and DCMI we can find both nouns and verbs.

4.2 Formalization and implementation

In this stage, we provide an explicit representation of the conceptualization captured in the previous stage in a formal language. The output of this life cycle phase was a *.owl* file, created using the ontology editor and framework for building intelligent systems called "Protégé Desktop 5.0". This file includes the formal definition of our conceptualization model. During the ontology implementation phase, we worked with the open source Java framework for Semantic Web and Linked Data applications "JENA"[13]. As a result, the Semantic Package version 1.1[14] ontology in a permanent link was obtained. A descriptive visual graph can

[6] `http://dublincore.org/`, last access Jun 2018
[7] `http://www.dbpedia.org/`, last access Jun 2018
[8] `https://w3id.org/nlp/semanticpackage/webpage`, last access Jun 2018
[9] `https://w3id.org/nlp/semanticpackage/1.1`, last access Jun 2018
[10] `https://w3id.org/nlp/semanticpackage/bucket_BarcelonaOpen2015`, last access Jun 2018
[11] `http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core`, last access Jun 2018
[12] `http://wordnet-rdf.princeton.edu/ontology`, last access Jun 2018
[13] `https://jena.apache.org/documentation/ontology/`, last access Jun 2018
[14] `https://w3id.org/nlp/semanticpackage/1.1`, last access Jun 2018

Source	Class concepts
Dbpedia	dbp:Person, dbp:Category, dbp:Place, dbp:Concept, dbp:Summary, dbp:Linguistics, dbp:Document, dbp:Miscellany, dbp:Named_entity, dbp:Organization, dbp:Semantic_class, dbp:Semantics, dbp:Lexis_(linguistics), dbp:Temporal_annotation, dbp:Polarity_(linguistics), dbp:Sentence_(linguistics), dbp:Taxonomy_(general)
SemanticPackage	sem_1_1:SUMO, sem_1_1:Semantic_Package, sem_1_1:Source, sem_1_1:Source_Type, sem_1_1:WNAffect, sem_1_1:WNDomain, sem_1_1:Lexical, sem_1_1:Sentiment_polarity, nif:Sentence, sem_1_1:Taxonomy
Wordnet Ontology	wordnet-ontology:Synset
Source	**Annotations**
Dbpedia	dbpedia:abstract, dbpediap:contactInfo dbpediap:copyright, dbp:Category
ONTOLegolang	ONTOLegolang_UAge:numberOfClasses, ONTOLegolang_UAge:numberOfDataProperties, ONTOLegolang_UAge:numberOfLogicalAxioms, ONTOLegolang_UAge:numberOfObjectProperties
Dublin Core	dc:creator, dc:date, purl:dateCopyrighted, purl:language, purl:license
Source	**Object Properties**
Dbpedia	dbpediap:sourceType
SemanticPackage	sem_1_1:conceptualized_by, sem_1_1:conceptualizes, sem_1_1:contained_by, sem_1_1:contains, sem_1_1:contains_Document, sem_1_1:contains_Sentence, sem_1_1:contains_Entity, sem_1_1:contains_synset, sem_1_1:direct_relation, sem_1_1:contains_Temporal_Info,sem_1_1:generated_from, sem_1_1:generates, sem_1_1:inverse_relation, sem_1_1:is_a
Dublin Core	purl:source
Source	**Data Properties**
Dbpedia	dbpediap:body, nif:lemma, dbpediap:offset, dbpediap:order, dbpediap:url
Dublin Core	purl:date, purl:hasVersion
Wordnet Ontology	wordnet-ontology:gloss
RDF Schema voc.	rdfs:label
	Prefixes
ONTOLegolang_UAge	`https://w3id.org/nlp/ONTOLegolang\_UAge`
dbp	`http://dbpedia.org/resource/`
dbpedia	`http://dbpedia.org/ontology/`
dbpediap	`http://dbpedia.org/property/`
purl	`http://purl.org/dc/terms/`
dc	`http://purl.org/dc/elements/1.1/`
xsd	`http://www.w3.org/2001/XMLSchema\#`
rdfs	`http://www.w3.org/2000/01/rdf-schema\#`
rdf	`http://www.w3.org/1999/02/22-rdf-syntax-ns\#`
wordnet-ontology	`http://wordnet-rdf.princeton.edu/ontology\#`
nif	`http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core\#`

Table 1: Reused terms

be watched by following this link: `http://visualdataweb.de/webvowl/#iri=https://w3id.org/nlp/semanticpackage/1.1`, last access Jun 2018.

4.3 Transition impacts

In this section we provide a detailed comparison between the original version 1.0 [15] of the Semantic Package and the current version 1.1. In this way, people that use the original version are able to identify key modifications for continuing

[15] `https://w3id.org/nlp/semanticpackage`, last access Jun 2018

Axiom	Description	BaseLine Axiom	Action
		Classes	
		Entity_Evaluation	Deleted
Linguistics		Linguistic	Name changed reused from DBpedia
		Sorted_Element	Deleted
		Sorted_Sentence	Deleted
	EquivalentTo Lexis from DBpedia	Lexical	
	EquivalentTo Sentence_(linguistics)	Sentence	
Temporal_annotation		Temporal_Information	Name changed to reuse Temporal_annotation from DBpedia, subclassOf Named_Entity
Semantics		Semantic	Name changed to reuse Semantics from DBpedia
Concept		Class	Name changed to reuse Concept from DBpedia
Semantic_class		Semantic_Class	Name changed to reuse Semantic_class from DBpedia
Sentiment_polarity	EquivalentTo Polarity_(linguistics), Name normalised according to the rest by using lower case in the second word	Sentiment_Polarity	
	EquivalentTo Sentence_(linguistics)	Sentence	
	EquivalentTo Taxonomy_(general)	Taxonomy	
WNAfects		Affects	Renamed for a better understanding
WNDomain		Domain	Renamed for a better understanding
		ObjectProperties	
Source_Type			New
contains_Document		document	Renamed
contains_Entity		entity	Renamed
contains_Sentence		sentence	Renamed
		sorted_Sentence	Deleted
contains_synset		synset	Renamed
contains_TemporalInfo		temporal_Information	Renamed
		DataProperties	
date	dateTime	Name changed to reuse date from Dublin core	
hasVersion	wordnet_version	Name changed to reuse hasVersion from Dublin core	

Table 2: Changes between the original version of Semantic Package, and the current version.

using this improved version. As it was previously mentioned, Table 2 shows the impacts to migrate to this version 1.1.

5 A case of study for capturing meaning from documents

The process of ontology population does not change the structure of an ontology, i.e., the concept hierarchy and non-taxonomic relations are not modified. What changes are the set of concept realizations (instances) and relations in the domain. Even, frequently, the schemes and the instance repositories appear in separate files, databases, etc. The most important element in this case is to ensure that the repositories are being built guided by appropriate schemes.

Most of the automatic ontology population processes involve at least one of the following strategies or a mixture of them[Petasis et al., 2011]:

- heuristics, in order to merge instances that refer to the same real object or event [Alani et al., 2003].
- special mapping rules, during instance creation (i.e. before the instances populate the ontology), in order to re-use instances that refer to the same real object or event instead of creating new ones [Buitelaar et al., 2006].

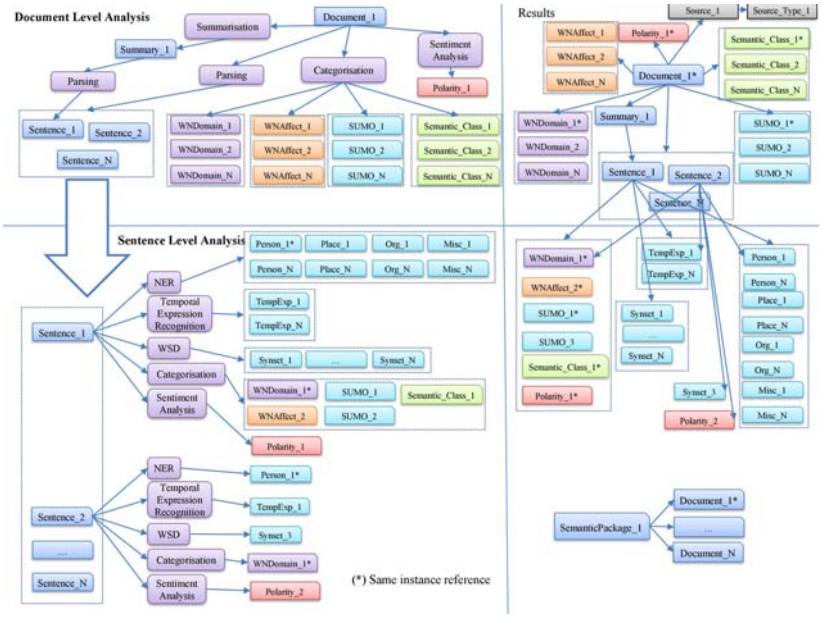

Figure 1: Process to extract and relate semantic data from documents.

HLT task	Tool name	Input and Output
Semantic Analysis [Gutiérrez et al., 2017], WN classes, WND relevant domains, WN-Affect, relevant SUMO categories	ISR-WN [Gutiérrez et al., 2011, Gutiérrez et al., 2016]	**Input:** Text (i.e. Documents, Sentences) **Output:** Disambiguated word senses, relevant semantic
Sentiment Analysis	Sentiment [Fernández et al., 2013]	**Input:** Text (i.e. Documents, Sentences) **Output:** Polarity (positive, negative, neutral)
Text Summarization	Compendium [Lloret and Palomar, 2012]	**Input:** Text (i.e. Documents, Sentences) **Output:** Most relevant sentences
Named Entity Recognizer	Stanford NER [Finkel and Manning, 2010]	**Input:** Text (i.e. Documents, Sentences) **Output:** Person, location, organization, and misc named entities
Temporal Expression Recognition	TipSem [Llorens et al., 2013]	**Input:** Text (i.e. Documents, Sentences) **Output:** Person, location, organization, and misc named entities

Table 3: HLT tools employed for identifying and extracting the instances for the ontology.

– the use of machine learning instead of manually-developed heuristics [Castano et al., 2009]

It can be said that ontology population systems are closely related to ontology-based information extraction systems, since the latter provide mechanisms to associate pieces of the data with concepts of an ontology. Thus, every ontology-based information extraction system can be viewed as an ontology population system, as it can be extended to assimilate extracted instances into the ontology [Petasis et al., 2011]. In our case, the process to automatically populate the ontology repository depends on the different HLT used and the schema designed. We use mapping rules, manually created, which depend on different heuristics

from outputs provided by HLT tools used for extracting knowledge form text. These rules link the HLT outputs we used to our ontology definition. No further techniques like machine learning, LSA (Latent Semantic Analysis) or Word2Vect were necessary to create them because the HLT output structure is usually fixed, as well as our ontology definition.

The process for capturing meaning presented in this research should follow an execution order to create instances and link them while the document is being processed (Figure 1). Each output next mentioned refers to a specific tool of Table 3. Moreover, the overall functioning of the algorithm is described in Algorithm 1.

Algorithm 1 Semantic Package population.

1: **procedure** OntologyPopulation
2: *input*:
3: $document \leftarrow$ the instance of a document in the ontology
4: $ont \leftarrow$ the ontology to populate
5: *output*:
6: $ont \leftarrow$ the populated ontology
7: *begin*:
8: $package \leftarrow$ **CreatePackageInstance**()
9: $\left.\begin{matrix} sentences[], wndomains[], wnaffects[], sumo\_classes[] \\ sem\_classes[], polarity, summary, sum\_sentences[] \end{matrix}\right\} \leftarrow$ **ProcessHlt**($document$)
10: ont.**AddInstances** $\left(\begin{matrix} package, document, wndomains, wnaffects, sumo\_classes, \\ sem\_classes, polarity, summary \end{matrix}\right)$
11: ont.**AddRelations** $\left(\begin{matrix} package, document, wndomains, wnaffects, sumo\_classes, \\ sem\_classes, polarity, summary \end{matrix}\right)$
12: ont.**AddRelation**($summary, sum\_sentences$)
13: **for** $sentence : sentences$ **do**
14: ont.**AddInstance**($sentence$)
15: ont.**AddRelation**($document, sentence$)
16: **ProcessSentence**($ont, document, sentence$)
17: **end for**
18: **end procedure**

The process starts by considering a **document** as input and creating a semantic package (line 8) which is the link between the different linguistic elements for a given document. Then, that document is processed in line 9 using different HLT tools for obtaining the following elements:

sentences A syntactic parser to split the document into sentences (**Parsing**).

WNDomains, WNAffects, SUMO and Semantic Class A semantic analyzer to obtain **categories** at document level (**Categorisation**).

polarity A sentiment analyzer to obtain the **sentiment polarity** at document level such as: Positive, Negative or Neutral (**Sentiment Analysis**).

summary A summarizer tool is able to reduce the document creating another document including only the most relevant sentences (**Summarisation**).

- The document obtained by the summarization process is also parsed into **sentences**. This allows to set semantic references between the sentences included into the main document and the sentences that form part of summary (**Parsing**).

Once all these semantic and lexical data at document level are obtained, all these data are indexed as singular instances (line 10) and the relation between them are created (lines 11 and 12). These instances will be used as semantic references defined in Table 4 from the sentence level analysis. It is important to highlight that the sentences obtained from the summary process are a subset of the sentences of the document ($sum\_sentences[] \subset sentences[]$). For this reason in the array $sum\_sentences[]$ only contains references to the array $sentences[]$ and all future process applied to the document sentences are also related with the summary sentences. Notice that each document should have associated information about its origin source and source type. This information is needed to identify and associate documents taking into account these data.

For each sentence of the document, an ontology instance is created and related with the original document (lines 14 and 15). Finally, each sentence is processed using also HLT tools (line 16) with a procedure defined in the Algorithm 2.

Algorithm 2 Sentence processing and results integration in the ontology.

```
1: procedure ProcessAndAddSentenceToOntology
2: input:
3:     ont ← the ontology to populate
4:     sentence ← an instance of a document sentence
5: output:
6:     ont ← the populated ontology
7: begin:
8:     entities[] ← StanfordNER(sentence)
9:     dates[] ← TipSem(sentence)
10:    wndomain ← WnDomainClassifier(sentence)
11:    wnaffect ← WnAffectClassifier(sentence)
12:    sumo ← SumoClassifier(sentence)
13:    semclass ← SemanticClassifier(sentence)
14:    pol ← SentimentAnalysis(sentence)
15:    ont.AddInstances(entities, dates, wndomain, wnaffect, sumo, semclass, polarity)
16:    ont.AddRelations(sentence, entities, dates, wndomain, wnaffect, sumo, semclass, pol)
17:    for term : sentence do
18:        synset ← WnDisambiguator(sentence, term)
19:        ont.AddInstances(term, synset)
20:        ont.AddRelations(sentence, term, synset)
21:    end for
22: end procedure
```

In this second procedure, the HLT tools are applied at sentence level (lines 8-14). As a result of this process we obtain:

– Named Entities using the Stanford NER tool in order to identify Persons, Places, Organizations or Miscellanies (non-classified entities) mentioned in the text (**NER**).

– Temporal recognition is applied to identify time expression in the text and set a date as output (e.g. next Monday, output considering the current day 27/07/2017 the date would be 31/07/2017). Note that this process sets as current date, the day in which is being processed the document. So, it is very important to process documents the same date they are posted, in real time, since in our case of study we use **TipSem** [Llorens et al., 2013], which set this characteristic.

– Once again a semantic categorization process is applied, but now at a sentence level in order to determine relevant categories: WNDomains, WNAffects, SUMO and Semantic Class (**Categorisation**).

– And, finally, a sentiment analysis process is applied to obtain the sentiment polarity at sentence level such as: positive, negative or neutral (**Sentiment**).

Having the output per sentence, the next step is to create new semantic instances of each output (line 15) by considering the classes designed in the ontology, removing possible duplicated instances: if an instance already exists then, the algorithm links them instead of creating a new one. On this way, the ontology is populated by means instances and their relations (line 16). As it can be seen in Figure 1, the results at sentence level are also linked to the sentences already instanced in the first process. Notice, the most valuable data in this process is to extract semantic and lexical information from documents by considering the semantic relation existing among them by reusing common elements. Finally, the algorithm obtains the most appropriated word sense of each sentence's term (lines from 17 to 21) being it added and related into the ontology repository. This disambiguation task is applied to identify the exact meaning of the words in the text. The output will be a list of word senses based on the WordNet resource (**WSD**). In this case, the word senses used for populating the ontology repository are linked to WordNet RDF online available in `http://wordnet-rdf.princeton.edu` , last access Jun 2018.

It is important to comment on the fact that our ontology schema proposes to make use of some semantic resources, which are represented in a taxonomic structure. This implies that if the category outputs provide categories which are easily identified as father or child (heritance, detected by using ISR-WN [Gutiérrez et al., 2011, Gutiérrez et al., 2016]), it is necessary to represent this information by using the relation *is_a* (see Table 4).

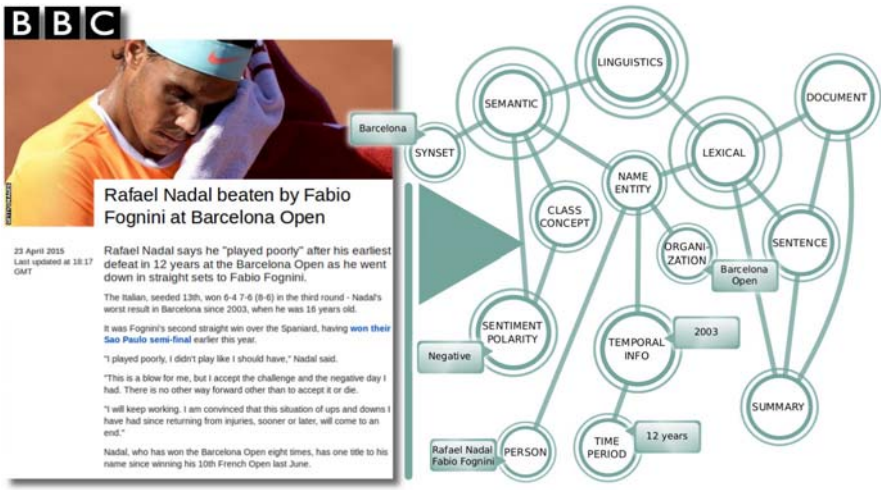

Figure 2: Case study of the semantic package ontology.

5.1 Example of ontology population

An example of the algorithm presented in the previous section for a case study is shown in Figure 2. This scenario is formed by a preliminary document, divided into two smaller subdocuments, reporting sport news, and more specifically, a news from a tennis match between Rafael Nadal and Fabio Fognini in the Barcelona Open 2015 competition extracted from the BBC news Website[16].

The reason why we selected this scenario for generating a testing repository (i.e. ontology instance repository), was due to the fact that this type of news is informative enough (it normally provides dates, named entities, key information of the match, etc.), to check whether our ontology proposal could capture all its semantics, or determine what important information could be missing, and therefore, improve the ontology in this respect. The ontology population was performed by following the process described in Section 5, using the tools described in Table 3, and obtaining some of the instances showed in Figure 3.

In this figure, the person "Rafael Nadal" is the main person, and in some sentences it appears related to the person "Fabio Fognini"; the organization "Barcelona Open", the domain concept "Tennis" among others. The negative judgment of *Sentiment polarity* is also present. With a deeper look in the phrase *sentence_p1_d2_s3*: "My forehand has been my biggest virtue, Nadal said; but my forehand was vulgar, it wasn't a forehand worthy of my ranking and career.", it can be appreciated how we are able to represent meta information described

[16] `https://www.bbc.com/sport/tennis/32436695`, last access Jun 2018

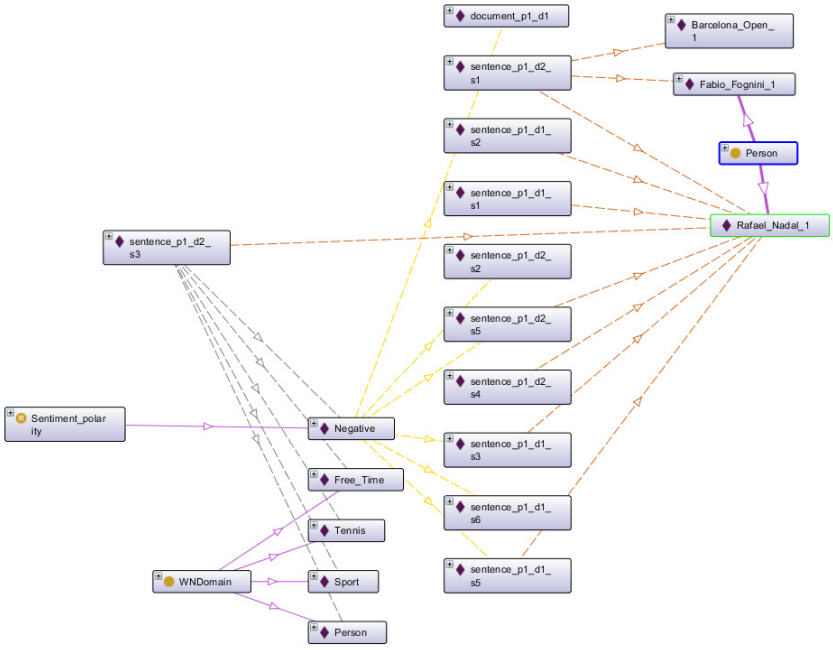

Figure 3: Example of instances in the ontology.

textually by means of semantic elements. So considering this potential, besides all semantic relationships represented among documents and sentences, this new vision of documents serves as a means of providing analytics over large textual information by using advanced queries, such as the ones presented in the next section.

Specifically, Stanford NER (Name Entity Recognition) tool is used to obtain the document structure as well as lexical information. Moreover, with this tool we are able to obtain name entities. All data extracted from the original text, serve to populating the ontology repository, building the relations between terms and the lexical information. As Figure 2 shows, it is possible to identify *Barcelona Open* as an organization, and *Rafael Nadal* and *Fabio Fognini* as person entities. These named entities bear a relationship with the sentences in which they appear by means of the semantic links described in Table 4.

Finally, the following tools are used for populating all the information that appears in Figure 2: TipSem to extract and standardize dates; ISR-WN for obtaining the relevant semantic categories of each sentence; Sentiment for classifying the sentence polarities, and so on. All this information is included and linked in the ontology repository, so, users can resolve some questions about the

Class	Relation	Ranges	Annot. & Card.	Inherited relationship from
dbp:Category	sem_1_1:conceptualizes	dbpedia:Document dbpedia:Summary nif:Sentence	NxN	-
dbp:Concept	-	-	-	-
dbp:Document	sem_1_1:conceptualized_by^{i}	dbp:Category	NxN	
	sem_1_1:contains_Sentence	nif:Sentence	dbp:order (NxN)	
	purl:source ^{f}	sem_1_1:Source	Nx1	
	sem_1_1:generates	dbp:Summary	NxN	
	sem_1_1:generated_from^{i}	dbp:Document	NxN	
sem_1_1:Lexical	-	-	-	-
dbp:Linguistics	-	-	-	-
dbp:Miscellany	-	-	-	-
dbp:Named_entity	-	-	-	-
dbp:Organization	-	-	-	-
dbp:Person	-	-	-	-
dbp:Place	-	-	-	-
sem_1_1:Sentiment_polarity	sem_1_1:conceptualizes	dbpedia:Document dbpedia:Summary nif:Sentence	NxN	dbp:Category
dbp:Semantic_class	sem_1_1:conceptualizes	dbpedia:Document dbpedia:Summary nif:Sentence	NxN	dbp:Category wordnet-ontology:Synset
dbp:Semantics	-	-	-	-
nif:Sentence	sem_1_1:conceptualized_by^{i}	dbp:Category	NxN	-
	sem_1_1:contains_Entity	dbp:Named_entity	NxN	-
	sem_1_1:contains_Synset	wordnet-ontology:Synset	NxN	-
	sem_1_1:contains_Temporal_Info	dbp:Temporal_annotation	NxN	-
dbp:Summary	sem_1_1:conceptualized_by^{i}	dbp:Category	NxN	dbp:Document
	sem_1_1:contains_Sentence	nif:Sentence	dbp:order (NxN)	dbp:Document
	purl:source	sem_1_1:Source	Nx1	dbp:Document
	sem_1_1:generates	dbp:Summary	NxN	dbp:Document
	sem_1_1:generated_from^{i}	dbp:Document	NxN	dbp:Document
wordnet-ontology:Synset	-	-	-	-
dbp:Taxonomy	sem_1_1:is_a^{r}	dbp:Taxonomy	NxN	-
dbp:Temporal_annotation	-	-	-	-
sem_1_1:Semantic_Package	sem_1_1:contains_Document	dbp:Document	NxN	-
sem_1_1:Source	dbpediap:sourceType	sem_1_1:Source_Type	NxN	-
-	sem_1_1:contains^{t}	-	-	-
-	sem_1_1:contained_by^{i,t}	-	-	-
sem_1_1:Source_Type	-	-	-	-
sem_1_1:SUMO	-	-	-	sem_1_1:Taxonomy
sem_1_1:WNAffect	-	-	-	sem_1_1:Taxonomy
sem_1_1:WNDomain	-	-	-	sem_1_1:Taxonomy

Table 4: Semantic relationships. Functional (f), Transitive(t), Reflexive(r), Inverse (i)

text, such as "*how many people are named in the text?*", "*in which summaries the named entity Rafael Nadal appears?*" or "*in which positive sentences Rafael Nadal is named in news talking about the Barcelona Open championship?*".

6 Evaluation and control

Evaluating an ontology means determining the quality of the final representation in terms of maintenance and reusability. The output of this phase, including both verification and validation calculations and results, is described next. The documentation output of this phase is also available online[17].

To guarantee the quality of our ontology proposal, a set of outputs obtained in the design and development life cycle are included. By quality, we understand the degree to which a set of functional and physical characteristics matches

[17] `https://w3id.org/nlp/semanticpackage/1.1/doc`, last access Jun 2018

the needs and expectations established in the specification phase [ISO, 2015]. Unfortunately, there is a disagreement on the way qualitative and quantitative validations are carried out [Yao et al., 2005, Blomqvist et al., 1989, Cross and Pal, 2008]. However, the current trend is to accept that the main purpose of an evaluation is to check that conceptualization model matches the adequacy of its content (validation) to determine their usefulness and potential for reusing. The aim of validating the ontology schema consists in ensuring the lack of construction errors or defects.

Verifying means ensuring that the ontology schema definitions match (as close as possible) the domain for which it was created.

6.1 Qualitative validation

Our task oriented ontology is according to requirements for representing different features identified in documents. It includes one main class (*Linguistics*), 25 subclasses and two isolated classes (*Source* and *Source_Type*). Each subclass may have additional properties, such that its parent class cannot be declared in its own level of generalization. Note that we included some classes used as equivalent of others for providing a better semantic support. Those are *dbo:Lexis_(linguistics)*, *dbo:Polarity_(linguistics)*, *dbo:Sentence_(linguistics)* and *dbo:Taxonomy_(general)*.

Our ontology schema does not contain any loop issues, with the exception of *Taxonomy*, in the hierarchical structure modeled (i.e., it does not have any class defined as a generalization and specialization of itself). In case of instancing *Taxonomy* the user should ensure a tree structure in its dynamic conceptual representation. The hierarchical relationship between subclasses is transitive (if B is a subclass of A and C is a subclass of B, then C is a subclass of A) and all the declared sibling classes in the hierarchy are at the same level of granularity (see class hierarchy of [Lloret et al., 2015]). Other transitive relationships, which serve to provide a better level of inference to our ontology schema, can be found in Table 4. Also, notice that a functional attribute, i.e. source, sets the cardinality for a singular value (N:1).

The subclasses are also related through non-hierarchical relationships. We have declared 13 different types of active relationships (as shown in Table 4, leaf object properties) with their respective cardinality (those checked as functional). Furthermore, this ontology schema has been twofold tested applying the standard validator w3c[18] and by means Jena[19] reasoning [20], both resulting successful. In this manner, style, format and redundancy issues are validated.

[18] `https://www.w3.org/RDF/Validator`, last access Jun 2018

[19] `https://jena.apache.org/documentation/ontology/`, last access Jun 2018

[20] When the reasoner is started up, in this case we used HermiT 1.3, this checks the ontology consistency and shows the possible errors that could appear.

6.2 Quantitative validation

To determine the physical characteristics of the structure and the type of content described in our ontology schema, we selected some of the metrics proposed in [Yao et al., 2005, Blomqvist et al., 1989, Cross and Pal, 2008]. Descriptive metrics show that our ontology is a small task-oriented ontology of a high-specialized domain (see first group of metrics in Table 5), with an appropriate and balanced weight in both vertical and horizontal axes of the inheritance tree (deduced by the result of 2.85 in the inheritance density parameter) [Tartir et al., 2005]. The relative low density of our ontology (average subclasses by concept 0.92) illustrates the restrictions mentioned in the previous subsection. These limitations also explain the cohesion achieved (average depth of inheritance 0.65), which is moderate, but very close to an average level [Yao et al., 2005, Blomqvist et al., 1989]. However, we can claim that the main advantage of our ontology lies in the completeness of relations and declared properties. In our ontology, non-taxonomic relationships density (averaged by concept 1.20) show its potential for inference [Cross and Pal, 2008] (relationship density 1.54), as well as for reusing it in other possible future goals (average of relationships reused by concept 0.16) [Cross and Pal, 2008]. Finally, we can also deduce its knowledge density, since our ontology is extensive and detailed (property density 1.04). This makes the population easier with either low or high density data, in a manual or automatic way [Cross and Pal, 2008].

With respect to the metrics provided in Table 5, it is important to clarify the following issues: for all these equations c represents the total number of concepts; the Equation (1) refers the sum of class axioms: SubClassOf counts; (2) refers the sum of object properties: Transitive, Inverse, Functional, SubPropertyOf, Symmetric, etc; in (3) $s(i)$ represents the number of subclasses of a concept i; in (4) $r(i)$ describes the total number of taxonomic relationships of a concept i; in (5) $r_{not}(i)$ represents the number of non-taxonomic relationships of a concept i; in (6) $reused_{rel}(i)$ describes the number of reused DCMI terms of a concept i; in (7) $reused_{prop}(i)$ represents the number of reused DCMI attributes of a concept i; in (8) $path(i)$ describes the deepest path from a concept i to a leaf node; in (9) $n_{att}(i)$ represents the number of data properties/attributes of a concept i and $n_{rel}(i)$ the number of object properties of a concept i; in (10) $sc(i)$ describes the number of subclasses of a concept i; in (11) $tax_{rel}(i)$ represents the number of taxonomic relationships of a concept i and $sem_{rel}(i)$ the number of non-taxonomic (semantic) relationships; (*) are minimum (min) and maximum (max) values.

6.3 Validating the competence questions

To verify that the ontology schema is able to extract the information for which it was designed and developed, a set of 30 competence questions was reused from

Metric	**Equation**	**Result**
Class Count	Protégé 5.0 pluging	28
Object Property Count	-	16
Data Property Count	-	8
Annotation Terms Count	-	14
External Reused Term	-	37
Root Concepts N.	-	1
Leaf Concepts N.	-	19
Taxonomic Relationships N.	Protégé 5.0 pluging (1)	23
Other non-Taxonomic rel.	Protégé 5.0 pluging (2)	7
Equivalent Relationships	-	4
Reused Classes	-	17
Reused ObjectProperties	-	2
Reused DataProperties	-	8
Average number of...		
(i)...subclasses (Avg.subclasses.n)	$\frac{\sum_{i=1}^{c} s(i)}{c}$ (3)	0.92 [min-0,max-5]*
(ii) ...taxonomic rel. by concept (Avg.rel.n)	$\frac{\sum_{i=1}^{c} r(i)}{c}$ (4)	1.00 [min-0,max-5]*
(iii) ...non-taxonomic rel. by concept (Avg.nonTrel.n)	$\frac{\sum_{i=1}^{c} r_{not}(i)}{c}$ (5)	1.20 [min-0,max-9]*
(iv) ...semantic reused rel. by concept (Avg.reuse rel)	$\frac{\sum_{i=1}^{c} reused_{rel}(i)}{c}$ (6)	0.16 [min-0,max-2]*
(v) ...reused attributes by concept (Avg.reuse prop)	$\frac{\sum_{i=1}^{c} reused_{prop}(i)}{c}$ (7)	0.40 [min-0,max-4]*
Avg. depth of inheritance by concept (Avg.depth)	$\frac{\sum_{i=1}^{c} max(path(i))}{c}$ (8)	0.65 [min-0,max-4]*
Property density (Prop.density)	$\frac{\sum_{i=1}^{c} n_{att}(i)+n_{rel}(i)}{c}$ (9)	1.04 [min-0,max-5]*
Inheritance density (Inh.density)	$\frac{\sum_{i=1}^{c} sc(i)}{c}$ (10)	2.85 [min-0,max-28]*
Relationship density (Rel.density)	$\frac{\sum_{i=1}^{c} tax_{rel}(i)+sem_{rel}(i)}{c}$ (11)	1.54 [min-0,max-5]*

Table 5: Ontology metrics and results.

the original ontology and fit to the current one. Its aim consists of determining whether the ontology could provide a correct response to these questions, thus validating its correctness. The competence questions had different degrees of difficulty, ranging from simple questions (e.g. *what PLACE named entities are in the documents?*) to more complicated ones (e.g. *which are the positive and negative sentences that talk about the sports domain?*), or even *which PERSON named entities appear in the relevant sentences of the document? (i.e., in its summary).* Moreover, they were defined taking into account the two type of users that could benefit from this ontology (data analyst and HLT expert). Our purpose here was to translate the competence questions in natural language into SPARQL questions to be executed in a ontology repository guided by our proposed schema and assess if this is able to provide a correct answer for each

question.

Table 6 shows two examples of questions in natural language, the user type to whom the query would be more appropriate, their SPARQL translation, and the result obtained after querying the ontology. These questions were tested on a repository of this ontology schema[21], which was automatically populated by processing digital documents extracted from news from a tennis match between *Rafael Nadal and Fabio Fognini in the Barcelona Open 2015* competition extracted from the BBC news Website[22]. This automatic populated repository has been twofold tested by applying the standard validator W3C and Jena Reasoner and using the same validation performed to the ontology in Section 6.1. For both aspects, the resulting tests were successful. Therefore, style, format and redundancy issues are again validated. In terms of ontology population quality, due to this population approach follows a set of rules and heuristics, described in Section 5, the quality of the information instanced depends on the accuracy of the technologies involved.

Concerning the results of the competence question evaluation by considering this scenario, we obtained that 96.6% of the them were correctly answered by the ontology (i.e., 29 out of 30), thus meaning that it is reliable enough for extracting personalized information depending on the users needs. There was only one question for which the information required was not represented in our ontology schema. This was related to the type of questions asking for the evaluation of an element at a global level, for instance, when one wants to ask which documents the entity X (e.g., Rafa Nadal) is positively and negatively considered. To be able to respond to this type of question, a change in the ontology design would be needed, as it is analyzed and discussed in Section 6.4.

Please note this repository can be queried online by using the competency questions formulated in Table 6 or others freely generated by users.

6.4 Discussion of the results

Although we showed that the ontology schema is able to capture and provide the information for which it was designed, from the analysis of the competence questions, we also realized that it may have limitations for a particular type of questions, as it was previously mentioned. In this respect, the ontology is not able to directly answer questions like *"what is the polarity for the entity X ?"* or *"which documents negatively refer to the entity Y ?"*. This is due to the fact that at this state we cannot capture multi-aspect polarity for the entities involved in a document, although we could obtain the sentences in which a specific entity is considered positive, negative or neutral and deduce the polarity of the entity from

[21] `https://w3id.org/nlp/semanticpackage/bucket_BarcelonaOpen2015`, last access Jun 2018

[22] `https://www.bbc.com/sport/tennis/32436695`, last access Jun 2018

Query: Could I know which other types of entities appear in the same sentences as the ones mentioning Rafa Nadal negatively?
User type: Data Analyst
SPARQL:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX purl: <http://purl.org/dc/terms/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/resource/>
PREFIX dbpediap: <http://dbpedia.org/property/>
PREFIX ONTOLegolang_UAge: <https://w3id.org/nlp/ONTOLegolang_UAge#>
PREFIX sem_1_1: <https://w3id.org/nlp/semanticpackage/1.1#>
PREFIX bucket_BarcelonaOpen2015:
    <https://w3id.org/nlp/semanticpackage/bucket_BarcelonaOpen2015#>
Select DISTINCT ?entityExtra ?type ?polarity ?body
WHERE { ?sentence rdf:type nif:Sentence; dbpediap:body ?body;
  sem_1_1:contains_Entity ?entity; sem_1_1:contains_Entity ?entityExtra.
  ?entityExtra rdf:type ?type. ?type ?p dbp:Named_entity.
  ?sentence sem_1_1:conceptualized_by ?polarity.
  ?polarity rdf:type sem_1_1:Sentiment_polarity.
  FILTER ( regex(str(?polarity), 'Negative') ).
  FILTER ( regex(str(?entity), 'Nadal') && (?entity != ?entityExtra))}
GROUP BY ?entityExtra ?type ?p ?polarity ?body ORDER BY ASC (?entity)
```

Result:

```
?entityExtra: Italian_1 ?type:Miscellany ?polarity:Negative
    ?body:"The Italian, seeded 13th."
?entityExtra:Barceona_1 ?type:Place ?polarity:Negative
    ?body:"The Italian, seeded 13th."
?entityExtra: Fabio_Fognini_1 ?type:Person ?polarity:Negative
    ?body:"Nadal battled back."
?entityExtra: Italian_1 ?type:Miscellany ?polarity:Negative
    ?body:"Nadal battled back."
```

Query: Which entities of the documents document_p1_d1 are not mentioned in the summary of this document?
User type: NLP expert
SPARQL:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX purl: <http://purl.org/dc/terms/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/resource/>
PREFIX dbpediap: <http://dbpedia.org/property/>
PREFIX ONTOLegolang_UAge: <https://w3id.org/nlp/ONTOLegolang_UAge#>
PREFIX sem_1_1: <https://w3id.org/nlp/semanticpackage/1.1#>
PREFIX bucket_BarcelonaOpen2015:
  <https://w3id.org/nlp/semanticpackage/bucket_BarcelonaOpen2015#>
Select distinct ?entity
WHERE { bucket_BarcelonaOpen2015:document_p1_d1 rdf:type dbp:Document;
  sem_1_1:contains_Sentence ?sentence. ?sentence ?rel ?entity.
  ?entity rdf:type ?named_entity. ?named_entity rdfs:subClassOf dbp:Named_entity.
  bucket_BarcelonaOpen2015:document_p1_d1 sem_1_1:generates ?summary
  MINUS { ?summary ?s dbp:Summary; sem_1_1:contains_Sentence ?sentenceS.
    ?sentenceS ?rel ?entity. ?entity rdf:type ?named_entityS.
    ?named_entityS rdfs:subClassOf dbp:Named_entity. }}
```

Result:

```
?entity: Barcelona_1, ?type: Place
?entity: Italian_1, ?type: Miscellany
?entity: Sao_Paulo_1, ?type: Place
?entity: Spaniard_1, ?type: Miscellany
?entity: 2003-01-01T00:00:00, ?type: Temporal_annotation
?entity: 2015-01-01T00:00:00, ?type: Temporal_annotation
```

Table 6: Example of competence questions for validating the ontology, their translation to SPARQL and the results obtained.

this information. To overcome this limitation, the initial ontology design should be slightly modified, introducing a new concept that would store the information regarding its evaluation (e.g., polarity evaluation). This concept should be at the top level of the ontology schema.

Another issue to remark concerns the concept *Sorted_Element*, which was deleted from the initial version of the ontology schema. This concept was originally defined to be able to store the position of the sentences in the summary with respect to the original document, but in this version 1.1 we propose to use the annotation axiom *dbpediap:order* for this aim. Therefore, we can relate a *Document* with a *Sentence* by means the relation *contains_Sentence* also including the *dbpediap:order* annotation. This issue was considered as future work in the development of the previous ontology schema (i.e. v1.0) and now it is solved. By reusing NIF we realized three terms were relevant: Sentence, lemma and Opinion. However, Opinion is a NIF class which represents opinion values between -1 and 1, despite this class is similar to our Sentiment_polarity class we could not used because we pretend use classifications , i.e. Positive, Negative, Neutral. As future works both NIF classes Opinion and Sentiment_polarity are going to be studied for being aligned.

Regarding the potentials of the Semantic Package ontology schema version 1.1, we would like to stress upon the fact that despite it is not a big or complex ontology, it is able to easily determine and infer information that can be personalized depending on the users needs. For instance, in our illustrative scenario, one may be interested in obtaining only information about the performance of "Rafa Nadal", whereas other user could be more interested in knowing what other facts also happened in that match. Moreover, information obtained from different sources could be also related and deduced using this ontology. For example, if more documents had been tested for our use case scenario, we could have obtained a series of facts and sentences all of them related to a specic entity, polarity, domain, etc. Note that the competence questions developed for this work are generic and respond adequately to the scenario selected about a "Rafa Nadal news report". In it, the specific entities involved act as variables inside SPARQL queries. In this manner, any other scenario can be used if linguistic elements such as document, sentence, named entity, temporal information (date references), words (considering word sense), identification of conceptualizations (semantic classes, sentiment polarity, emotions - WNAffects, WNDomain, SUMO categories), and so on can be found.

One of the advantages of our proposed ontology schema is that, differently from other existing ontologies, this is a task-oriented ontology schema that captures the semantic of documents. Given that, this information can be obtained independently by different HLT tools, all these outputs can be integrated in a single-ontology to maximize the exploitation and allow better reasoning pro-

cesses. Since our medium-term goal is that any ontology repository based on Semantic Package version 1.1 could be also automatically populated from the output of these HLT tools, the ontology will then have another added-value, allowing that both humans or automatic processes can use the information contained to easily obtain and generate the type of information more suitable to their interests.

7 Conclusion and future work

This research proposed the redesign, development and validation of an ontology schema for representing textual information based on the use of Human Language Technologies tools, as well as a novel approach for automatically creating instance of documents to populate an ontology repository. The proposed ontology schema was qualitative and quantitative validated. Moreover, it was shown to be useful and correct, based on a comprehensive analysis and validation over a set of 30 competence questions. The results obtained showed that all the questions, except one were correctly answered. As result of this work we have significantly improved an existent ontology by considering shared schemas of the Semantic Web. In addition, we have provided three permanent online accesses: a schema for representing semantics of documents, its Web page documentation and a repository that stores semantic individuals of a real scenario.

Both the ontology schema definition and the approach to be automatically populate its repositories have great potential for tasks, such as natural language generation, since it could be exploited for generating personalized information, adapting the type of information to the users' or information needs. In the future, we would like to use the top performing HLT tools for extracting information from documents as well as to apply the automatic population approach with a large collection of heterogeneous texts, including those belonging to newswire, blogs, reviews, and tweets, among others. This manner we could analyze and obtain common information across different genres that would be later used for generating new texts. In addition, for a second stage of our research we plan to consider natural language processing features presented in lemon and NIF, which go deeper in terms of language's representation.

Acknowledgements

This research work has been partially funded by the University of Alicante, Generalitat Valenciana, Spanish Government, Ministerio de Educación, Cultura y Deporte and Ayudas Fundación BBVA a equipos de investigación científica 2016 through the projects TIN2015-65100-R, TIN2015-65136-C2-2-R, PROMETEU/2018/089,"Plataforma inteligente para recuperación, análisis y representación de la información generada por usuarios en Internet" (GRE16-01) and

"Anlisis de Sentimientos Aplicado a la Prevención del Suicidio en las Redes Sociales" (ASAP).

References

[Alani et al., 2003] Alani, H., Kim, S., Millard, D. E., Weal, M. J., Lewis, P. H., Hall, W., Shaboldt, N., and Shadbolt, N. (2003). Automatic Extraction of Knowledge from Web Documents. *Web and Web Services*, pages 1–11.

[Basile et al., 2016] Basile, V., Cabrio, E., and Schon, C. (2016). KNEWS: Using Logical and Lexical Semantics to Extract Knowledge from Natural Language. In *Proceedings of the European Conference on Artificial Intelligence (ECAI) 2016 conference.*

[Blomqvist et al., 1989] Blomqvist, E., Öhgren, A., and Sandkuhl, K. (1989). Ontology Construction in an Enterprise Context: Comparing and Evaluating two Approaches. *FLORIDA INTERNATIONAL UNIVERSITY. PRIOR.*

[Buitelaar et al., 2006] Buitelaar, P., Cimiano, P., Racioppa, S., and Siegel, M. (2006). Ontology-based information extraction with soba. *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 2321–2324.

[Castano et al., 2009] Castano, S., Peraldi, I. S. E., Ferrara, A., Karkaletsis, V., Kaya, A., Mller, R., Montanelli, S., Petasis, G., and Wessel, M. (2009). Multimedia interpretation for dynamic ontology evolution. In *Journal of Logic and Computation*, volume 19, pages 859–897.

[Celjuska and Vargas-vera, 2004] Celjuska, D. and Vargas-vera, D. M. (2004). Ontosophie: A Semi-Automatic System for Ontology Population from Text. In *In: International Conference on Natural Language Processing (ICON).*

[Chaves et al., 2012] Chaves, M. S., de Freitas, L. A., and Vieira, R. (2012). Hontology: A Multilingual Ontology for the Accommodation Sector in the Tourism Industry. In *KEOD 2012 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Barcelona, Spain, 4 - 7 October, 2012*, pages 149–154.

[Corcoglioniti et al., 2016] Corcoglioniti, F., Rospocher, M., and Aprosio, A. P. (2016). Frame-based ontology population with PIKES. *IEEE Trans. Knowl. Data Eng.*, 28(12):3261–3275.

[Costa et al., 2013] Costa, R., Figueiras, P., Maló, P. M. N., and Lima, C. (2013). Classification of Knowledge Representations using an Ontology-based Approach. In *KEOD 2013 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Vilamoura, Algarve, Portugal, 19-22 September, 2013*, pages 184–191.

[Cross and Pal, 2008] Cross, V. and Pal, A. (2008). An ontology analysis tool. *International Journal of General Systems*, 37(1):17–44.

[Draicchio et al., 2013] Draicchio, F., Gangemi, A., Presutti, V., and Nuzzolese, A. G. (2013). Fred: From natural language text to rdf and owl in one click. In Cimiano, P., Fernndez, M., Lopez, V., Schlobach, S., and Vlker, J., editors, *ESWC (Satellite Events)*, volume 7955 of *Lecture Notes in Computer Science*, pages 263–267. Springer.

[Faria et al., 2014] Faria, C., Serra, I., and Girardi, R. (2014). A domain-independent process for automatic ontology population from text. *Science of Computer Programming*, 95, Part 1:26–43.

[Fernández et al., 2013] Fernández, J., Gutiérrez, Y., Gómez, J. M., Martínez-Barco, P., Montoyo, A., and Munoz, R. (2013). Sentiment Analysis of Spanish Tweets Using a Ranking Algorithm and Skipgrams. *Proc. of the TASS workshop at XXIX Conference of the Spanish Society for Natural Language Processing (SEPLN 2013)*, pages 133–142.

[Finkel and Manning, 2010] Finkel, J. R. and Manning, C. D. (2010). Hierarchical Joint Learning: Improving Joint Parsing and Named Entity Recognition with Non-Jointly Labeled Data. In *Proceedings of ACL 2010.*

[Gruber, 1995] Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(5-6):907–928.

[Gutiérrez et al., 2010] Gutiérrez, Y., Fernández, A., Montoyo, A., and Vázquez, S. (2010). UMCC-DLSI: Integrative resource for disambiguation task. In *SemEval '10 Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 427–432. Association for Computational Linguistics.

[Gutiérrez et al., 2011] Gutiérrez, Y., Orquín, A. F., Vázquez, S., Montoyo, A., Gutiérrez Vázquez, Y., Fernández Orquín, A., Montoyo Guijarro, A., and Vázquez Pérez, S. (2011). Enriching the integration of semantic resources based on WordNet. *Procesamiento del Lenguaje Natural*, 47:249–257.

[Gutiérrez et al., 2016] Gutiérrez, Y., Vázquez, S., and Montoyo, A. (2016). A semantic framework for textual data enrichment. *Expert Systems with Applications*, 57:248–269.

[Gutiérrez et al., 2017] Gutiérrez, Y., Vázquez, S., and Montoyo, A. (2017). Spreading semantic information by word sense disambiguation. *Knowledge-Based Systems*, pages –.

[Hellmann et al., 2013] Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating NLP using linked data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8219 LNCS, pages 98–113.

[IBM, 1990] IBM (1990). Introduction to business system development method. Technical report, IBM.

[ISO, 2015] ISO (2015). ISO 9000:2015(en): Quality management systems. Fundamentals and vocabulary.

[KBSI, 1994] KBSI (1994). Knowledge based systems incorporated. Technical report, Wright-Patterson Air Force Base, Ohio.

[Krieger et al., 2012] Krieger, H.-U., Declerck, T., and Nedunchezhian, A. K. (2012). MFO-The Federated Financial Ontology for the MONNET Project. In *KEOD 2012 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Barcelona, Spain, 4 - 7 October, 2012.*, pages 327–330.

[Lehmann et al., 2012] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2012). DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia — www.semantic-web-journal.net. *Semantic Web Interoperability, Usability, Applicability an IOS Press Journal Search form*, 1(5):1–29.

[Llorens et al., 2013] Llorens, H., Saquete, E., and Navarro-Colorado, B. (2013). Applying semantic knowledge to the automatic processing of temporal expressions and events in natural language. *Information Processing & Management*, 49(1):179–197.

[Lloret et al., 2015] Lloret, E., Gutiérrez, Y., and Gómez, J. M. (2015). Developing an Ontology to Capture Documents' Semantics. In Fred, A. L. N., Dietz, J. L. G., Aveiro, D., Liu, K., and Filipe, J., editors, *KEOD 2015 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K) 2015), Volume 2, Lis*, pages 155–162. SciTePress.

[Lloret and Palomar, 2012] Lloret, E. and Palomar, M. (2012). COMPENDIUM: a text summarisation tool for generating summaries of multiple purposes, domains, and genres. *Natural Language Engineering*, 19(02):147–186.

[Petasis et al., 2011] Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., and Zavitsanos, E. (2011). Ontology population and enrichment: State of the art. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6050:134–166.

[Rospocher et al., 2016] Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T., and Bogaard, T. (2016). Building event-centric

knowledge graphs from news. *J. Web Sem.*, 37-38:132–151.

[Shen et al., 2012] Shen, W., Wang, J., Luo, P., and Wang, M. (2012). A Graph-based Approach for Ontology Population with Named Entities. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 345–354, New York, NY, USA. ACM.

[Sirin and Parsia, 2007] Sirin, E. and Parsia, B. (2007). SPARQL-DL: SPARQL query for OWL-DL. In *CEUR Workshop Proceedings*, volume 258.

[Suca and da Silva, 2013] Suca, E. G. and da Silva, F. S. C. (2013). An Ontology for Portability and Interoperability Digital Documents - An Approach in Document Engineering using Ontologies. In *KDIR/KMIS 2013 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval and the International Conference on Knowledge Management and Information Sharing, Vilamoura, Algarve, Portugal, 19 - 22 September, 2013*, pages 373–380.

[Tansley and Hayball, 1993] Tansley, S. and Hayball, C. C. (1993). Knowledge-based systems analysis and design - a KADS developer's handbook. pages I–XV, 1–528.

[Tartir et al., 2005] Tartir, S., Arpinar, B., Moore, M., Sheth, A., and Aleman-Meza, B. (2005). OntoQA: Metric-Based Ontology Quality Analysis. In *Proceedings of IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources*.

[Uschold and King, 1995] Uschold, M. and King, M. (1995). Towards a methodology for building ontologies. In *Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95*.

[Varile and Zampolli, 1997] Varile, G. B. and Zampolli, A. (1997). *Survey of the state of the art in human language technology*. Cambridge University Press.

[Yao et al., 2005] Yao, H., Orme, A. M., and Etzkorn, L. (2005). Cohesion Metrics for Ontology Design and Application. *Journal of Computer Science*, 1(1):107–113.