Revista de Lenguas para Fines Específicos 23.2 (2017), pp. 63-85
http://dx.doi.org/10.20420/rlfe.2017.173
ISSN: 2340-8561

Servicio de Publicaciones y Difusión Científica
Universidad de Las Palmas de Gran Canaria

Automatic domain-specific learning: towards a methodology for ontology enrichment

Pedro Ureña Gómez-Moreno[1]
University of Granada
Eva M. Mestre-Mestre
Universitat Politècnica de València

ABSTRACT

At the current rate of technological development, in a world where enormous amount of data are constantly created and in which the Internet is used as the primary means for information exchange, there exists a need for tools that help processing, analyzing and using that information. However, while the growth of information poses many opportunities for social and scientific advance, it has also highlighted the difficulties of extracting meaningful patterns from massive data. Ontologies have been claimed to play a major role in the processing of large-scale data, as they serve as universal models of knowledge representation, and are being studied as possible solutions to this. This paper presents a method for the automatic expansion of ontologies based on corpus and terminological data exploitation. The proposed "ontology enrichment method" (OEM) consists of a sequence of tasks aimed at classifying an input keyword automatically under its corresponding node within a target ontology. Results prove that the method can be successfully applied for the automatic classification of specialized units into a reference ontology.

Keywords: Ontology learning, FunGramKB, Corpus, Terminology, Biology

1. Introduction

Modern society is immersed in a process of rapid technological development in which data is constantly being generated and in which the Internet is used as the

[1] **Corresponding author** – Facultad de Ciencias de la Educación, Departamento de Didáctica de la Lengua y la Literatura, Campus Universitario de Cartuja, s/n, 18071, Granada, Spain.

Email: pedrou@ugr.es

primary means for information exchange. While the exponential growth of information poses many opportunities for social and scientific advance, it has also brought with it the so-called “knowledge acquisition bottleneck”, i.e. the inability of many current computational systems to extract meaningful patterns from massive data. Among the various solutions that have been proposed to solve this problem, ontologies have been claimed to play a major role in the processing of large-scale data, as they serve as universal models of knowledge representation. Their main advantage is that they allow sharing knowledge among diverse linguistic communities with little or no ambiguity. However, building ontologies is both time-consuming and labor-intensive, and it is thus necessary to create new systems which acquire knowledge more efficiently.

This paper aims to present a proof-of-concept process for ontology expansion in knowledge domains. The preliminary findings show that, even with a limited number of data, it can be effectively applied for the automatic classification of specialized units within reference ontologies. Its main advantage is that it draws on corpus and terminological data available from academic and encyclopedic sources.

The “ontology enrichment method” (OEM), as we have termed it, offers new opportunities for research in the processing of natural language. It consists of three main tasks: ontology identification, corpus compilation and automatic data classification. Based on it, this paper reports on the results of a small-scale experiment carried out in the field of virology to corroborate whether the OEM was able to classify an input keyword under its corresponding superordinate in a hierarchy of viruses.

After this brief introduction, Section 2 presents the state of the art, including a classification of ontology learning processes, and ontology extension and refinement processes. In Section 3, the proposal for automatic expansion of ontologies based on corpus and terminological data exploitation is presented. Section 4 describes the materials and methods used for the experiment with the OEM, and Section 5 details the results obtained. Finally, Sections 6 and 7 show the conclusions and discussion derived from the research.

2. State of the art

Ontology learning and expansion are key issues under focus in nowadays research towards a more complete and more effective Semantic Web. Indeed, due to the amount of data that needs to be analyzed and managed, better and improved ways of dealing with information and information systems are necessary. The incipient moment of semi-automatic processes seems to be long gone, as the needs of the linguist and the researcher call for entirely automatic processes able to solve the problem for ontology learning and expansion, by completing full procedures without the need for human intervention. In the ever-growing universe of data in modern information systems, and in order to make of the Semantic Web a reality, it becomes imperious that information should be mined, treated, analyzed, processed and used without the need for human interaction. In this context, ontologies are seen as a tool to tackle with data and semantics.

As the need for ontologies increased, the definition of what an ontology is, and how it has to be conceived and built, has also expanded and it has developed to include the different aspects of the process they entail, getting to include also the procedures of ontology learning, refinement and extension. Thus, according to Gruber (1993), ontologies are formal and explicit specifications structured as concepts and relations of shared conceptualizations, in the sense that they are commonly known and accepted as such. A later definition (Studer et al., 1998) explicated this characterization by saying that an ontology is a machine understandable description of clearly defined terms in which a fact is described in a commonly accepted abstract form. Wong et al. (2012) described them from a semantic perspective and explained that ontologies represent the intensional aspect of a domain to rule knowledge (i.e. extensional aspect). Regarding the ontology learning process, Gómez-Pérez & Manzano-Macho (2004) explained that it is the application of a series of techniques to build, and then enlarge, adapt and improve it, using heterogeneous knowledge and information sources.

2.1. Classification

First of all, we will look into some of the existing approaches to complete ontology learning processes. Thus, in order to construct, fill and complete an ontology, different types of sources can be used. These can help establish a first approach for a classification. Thus, an ontology can be enlarged with the learning of new ontological units by means of unstructured sources (i.e. using corpora created for this purpose, and then applying Natural Language Processing techniques to the texts), semi-structured sources (i.e. corpora created from text documents which have a previous structure, such as xml schemas), or structured sources, such as already existing databases or catalogues.

With regard to ontology learning, another approach that can be applied to attempt a classification is to look at the techniques used for the categorization of the elements included in the ontology. Four methods will be explained here: a) linguistics-based, b) statistics-based, c) logic-based or d) based on machine-learning methods. The first group is constituted by approaches based on linguistic techniques. They are often based on particular morphological or syntactic features existing in the texts used as corpora, together with the distributional information of specific elements under study e.g. collocations. They use different tools for their analyses, as will be explained below. The following can be mentioned: POS (Part of Speech) based patterns, phrase-based patterns, semantic lexicons, lexico-syntactic patterns, semantic templates, sub-categorization frames or seed words, and are usually dependent on natural language processing tools. To mention some, Gupta et al. (2002) and Haase & Stojanovic (2005) used part of speech tagging, and some examples of taggers are Brill Tagger (Brill, 1992) or Principar (Lin, 1994). And regarding sentence parsing, we will mention the proposal by Poon & Domingos (2010), and the Stanford Parser (Klein & Manning, 2003).

Other well-known linguistic techniques for ontology learning are syntactic structure analysis, which looks into syntactic information to expose relations at the sentence level; here, the head-modifier principle is used to identify hyponymy relations (Hippisley et al., 2005). Another line is dependency analysis (Gamallo et al., 2002; Ciaramita et al., 2005), in which grammatical associations are used to establish more complex relations. Some of the techniques based on semantic lexicons, which offer access to large collection of predefined concepts and relations, are related to the areas of lexical acquisition (MacNamara, 1982; O'Hara et al., 1998), word sense disambiguation (Ide & Véronis, 1997; Dorr & Jones, 1996), or similarity measurement (Pedersen et al., 2004). Alfonseca & Manandhar (2002) proposed a new linguistic approach, based on the Distributional Semantics Hypothesis, which looks for co-occurrences for each set of concepts, which can be used either to group concepts inside an ontology or to refine one with new concepts.

Focusing on statistics-based techniques, they do not take into consideration the linguistic component of the corpora. Here we are talking about techniques based on clustering, latent semantic analysis, co-occurrence analysis, term subsumption, contrastive analysis or association rule mining. Also, frequency analysis of word or pattern repetition and TF-IDF are usually applied to grade the relevance of the elements selected for analysis (Salton et al., 1975; Salton & Buckley, 1988).

One of the most commonly used approaches is clustering, seeking to group and establish a hierarchy for terms based on their similarity, centered on either individual terms or concepts (agglomerative clustering), or terms taken in bulk and then dividing them into groups (divisive clustering). For instance, Khan and Luo (2002) described a method for ontology construction by using a modified Self-Organization Map clustering algorithm in a bottom-up fashion. Hotho et al. (2001) and Lee et al. (2007) proposed various clustering techniques to view text documents with the help of an ontology.

Another approach is co-occurrence analysis, which aims to identify lexical units that appear together in the texts. Co-occurrence analysis limits the influence of popular terms not related to the domain (Roussinov & Zhao, 2003). For instance, Liu et al. (2005) developed a method based on WordNet (Princeton, 2010), whereas the approach used by Widdows et al. (2002) consisted in searching co-occurrence in lists of objects.

Moving on to logic-based analyses, they are the least common in ontology learning. They are used to deal with relations and axioms. There are two basic approaches in this perspective: inductive logic programming and logical inference. Inductive logic programming (ILP) (Lima et al., 2014; Lisi, 2005) induces symbolic extraction rules to populate a domain ontology with instances of entity classes by using domain-independent linguistic patterns. Concepts are divided into positive and negative examples. The characteristic feature of ILP, as compared to other forms of concept learning, is the use of prior knowledge during the induction process. Logical inference (Shamsfard & Barforoush, 2004; Udrea & Getoor, 2007) derives implicit relations from existing ones, using axioms such as inheritance and transitivity rules, and working first direct relations and, at a second level, indirect relations.

Finally, machine learning methods use algorithms to automatically analyze an ontology and establish relations among its elements, thus being able to establish their right place in the ontology assembly. There exist numerous examples of these techniques (cf. Mitchell, 1997; Hwang, 1999; Khan & Luo, 2002; Gacitua et al., 2008). The processes used in this approach usually involve an initial input provided by the researchers/users followed by the processing of lexical units in order to establish relations among them (e.g. “is-a” or “part of”), thus being able to create a hierarchy. The final step is to assign a lexical unit to a concept. In the process of creating the ontology, there exists a moment of concept learning, in which the machine learns how to identify concepts, and relations, so that it can be automatically enriched later on. Examples of this are for instance Missikoff et al. (2002), who elaborated an integrated approach, which can construct and access the domain ontology to integrate information intelligently within a virtual user community, or Navigli et al. (2004), who worked out the problem of semantic disambiguation based on WordNet and SemCor (Biemann, 2005). Evidently, there also exist hybrid approaches, which combine two or more techniques.

Wong et al. (2012) carried out a thorough analysis of existing methods for ontology learning and improvement, taking into account previous studies which looked into more than 80 approaches. The conclusions mentioned in these analyses highlighted the non-existence of a detailed methodology to guide ontology learning from text, as well as the lack of a completely automatic system for ontology learning, and pointed to the need for an approach to assess the accuracy of the process by comparing different systems. From a practical perspective, there was insufficient attention dedicated to non-taxonomic relations as compared to taxonomic relations, there is even less work dedicated to axiom learning, most of the ontology learning systems are domain-specific, without the possibility of cross-referencing, utilizing only one domain, and, finally, there is no standard method to assess the ontology learning process. Other studies pointed out that input data are mostly structured, and that the task of discovering relations is very complex, being the main hindrance to the progression of ontology learning. Among their conclusions, Wong et al. (2012) claimed that most existing ontologies are not complete, and this is a problem for the improvement of the ontology learning process and asserted the need to represent ontological entities as language-independent constructs. Also, the importance of ontology mapping to tackle with the significant amount of ontologies that are being created at present is mentioned.

2.2. Ontology extension and refinement

Ontology extension and refinement refer to slightly different processes, since they do not so much relate to the process of learning from ontologies, as the processes described above do, but aim to the improvement of language modelling from existing models. That is to say, based on already existing structures, they try to develop better models for language insertion in ontologies. As happened with ontology learning processes, and depending on the methods used, processes can be divided into automated and semi-automated approaches.

On the one hand, there exist some approaches that require the intervention of the expert at a given point in the process. Most of the existing methods (Faatz & Steinmetz, 2002; Liu et al., 2005; Biemann, 2005) are semi-automatic approaches. They are recall-based, that is, they usually work with statistically weighted terms, after comparing a given domain-specific corpus and common language databases. In some cases, they use co-occurrences in their analyses. On the other hand, the few existing automatic approaches (Hahn & Schnattinger, 1998) demand for a principled way to integrate new terms in the ontology, and aim to find validated rules for automatic expansion. Some domain-approached methods are for instance Navigli & Velardi (2002) in the realm of tourism or Lee et al. (2006) in the biomedical domain.

OntoLearn (Navigli & Velardi, 2004) starts with an existing generic ontology and a set of documents in a given domain, and produces a domain extended and trimmed version of the initial ontology. The ontology generated by OntoLearn is a linguistic ontology, since it is anchored to texts. It has been applied to different domains (e.g. tourism and computer networks), and consists of three phases: terminology extraction, semantic interpretation, and extending and trimming. A limitation of this approach is the occasional lack of focus when providing a definition for compound nouns, for instance, which affects the organization of concepts in hierarchies.

PACTOLE (Bendaoud et al., 2008) enriches an existing ontology by way of adding new texts and it uses experts to validate all the procedure in a five-step process. The first step is the extraction of domain terms and their properties. In the second step, a concept frame is built from the pairs object-property. In the third step the existing knowledge resources are converted into a lattice structure. Then, frames are merged. Finally, the result is represented with a description-logics formalism. Among the limitations of this approach, we can mention that the number of properties associated with objects might not be sufficient and, therefore, the resulting information is too small for classification; moreover, verbs are not the sole properties for defining a class, thus reducing the results; finally, some properties cannot be extracted by analyzers, due to their inherent nature.

3. Towards a methodology for ontology enrichment

This section presents a method for the automatic expansion of ontologies which is based on corpus and terminological data exploitation. As explained below, the proposed "ontology enrichment method" (OEM) consists of a sequence of tasks aimed at classifying an input keyword automatically under its corresponding node within a target ontology. In its current phase of development, the OEM is designed for the enrichment of "IS-A ontologies", i.e. taxonomies consisting of superordinate (or hyperordinate) concepts each of which subsumes one or more subordinate concepts. Additionally, the system is intended for enhancing both ontologies from specialized domains and general-purpose ontologies representing commonsense knowledge such as FunGramKB Core Ontology (Periñán-Pascual & Arcas, 2010). For illustrative purposes, the paper will focus on the former type.

A "concept" is defined in this paper as the minimum unit of knowledge representing any entity, event or quality in the real world or in an imaginary world. At the linguistic level, concepts are instantiated by lexical units, so that a conceptual unit can be linked to one or more linguistic expression(s). Similarly, this two-level schema is found in well-known databases such as WordNet, where the semantic unit called "synset" is lexically realized by networks of words. For example, the synset expressing "a motor vehicle with four wheels; usually propelled by an internal combustion engine" is instantiated by words such as "car", "auto", "machine" or "motorcar" at the lexical level (Princeton, 2010). Likewise, a distinction needs to be drawn between lexical attributes and ontological attributes. The former refers to any defining characteristic of words that may or may not be shared in all languages, whereas the latter refers to the inherent qualities of concepts, which are mental representations universally shared by every human. This distinction is relevant for the present paper insofar as the lexical attributes studied in the experiment are not conceptual in nature, since, as explained below, they were retrieved from linguistic sources. While adhering to this distinction – and for methodological purposes – the attributes of the ontology and the keyword will be both treated as conceptual in this paper.

The main assumption of the OEM is that the subordinate concepts that belong to a common superordinate in an ontology will necessarily share a set of common defining features or "attributes" which will identify them as members of the same class. In other words, if the semantic features of both the superordinates and subordinates can be found in advance, it is hypothetically possible to classify a new keyword by linking it automatically to a concept whose superordinate shows the strongest semantic affinity with the features of the keyword. For example, the concept "mammal" heads various subordinate concepts such as "wolf", "cow" or "squirrel" which, despite intracategorial differences, all share "having mammary glands for feeding" as a common defining feature. The question is then how to identify the set of relevant features among the conceptual units involved and how to use them for classification purposes. More formally, the task is defined as follows: "let *O* be an existing ontology which contains a set of superordinate concepts *H = {h1, h2, … hn}*, i.e. units that are not subsumed by higher-rank superordinates, and a set of subordinate concepts *S = {s1, s2, … sn}*, i.e. terminal units that do not subsume other lower-level nodes. Then define a function "f(h) := x →y" such that *x* is an input subordinate concept, which the input keyword is linked to, and *y* is the corresponding host superordinate (*y ∊ H*)".

Before the fundamentals of the OEM are explained in detail, some terminological notes are in order. While "term" is frequently found in the literature as an equivalent expression to "keyword", in this paper the former will be used to refer to any word or phrase that instantiates a concept from an expert domain such as physics, medicine or nutrition; on the other hand, "keyword" will be used to refer to an ngram, i.e. a group of characters between spaces.

4. Materials and method

The OEM consists of three main tasks: ontology identification, corpus compilation and data classification. Based on this methodology, this paper reports on the results of a small-scale experiment carried out in the field of virology and whose purpose was to examine whether the OEM was able to classify the input keyword *dengue* under its corresponding superordinate in a hierarchy of viruses.

4.1. Task 1: Ontology identification

The method starts with a pre-established ontology, which has been either manually created by experts or by means of ontology-induction tools. The ontology serves two main purposes. First, it must provide background lexicographic and/or encyclopedic information for each superordinate concept, including a description of the main events, qualities and related entities. These descriptors will be used as conceptual landmarks guiding the system in the process of accommodating the input keyword. Second, it represents the "gold standard" against which the precision and recall of the enrichment model is evaluated. In this regard, the OEM must classify an input keyword under the same conceptual node shown in the initial ontology.

For the experiment, the Baltimore virus classification was chosen as a reference ontology. This distinguishes seven major (i.e. superordinate) groups of viruses, ranging from "I" to "VII". Additionally, each group differentiates various viruses or genera (i.e. subordinates). Table 1 shows the main structure of the classification:

Group	Name	Example
I	double-stranded DNA viruses	adenoviruses
II	single-stranded DNA viruses	parvoviruses
III	double-stranded RNA viruses	reoviruses
IV	single-stranded RNA (+) viruses	picornaviruses
V	single-stranded RNA (-) viruses	orthomyxoviruses
VI	single-stranded RNA (RT) viruses	retroviruses
VII	double-stranded DNA (RT) viruses	hepadnaviruses

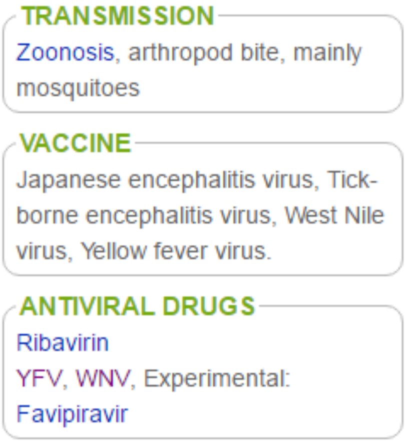
Table 1. Baltimore virus classification (adapted from Baltimore, 1971).

The *dengue* virus, which is a species of type IV, was selected as the keyword to be classified. We also selected at least one instance virus from each class, namely *aviadenovirus* (group I), *circovirus* (group II), *phytoreovirus* (group III), *flavivirus* (group IV), *hepacivirus* (group IV), *henipavirus* (group V), *deltaretrovirus* (group VI) and *orthohepadnavirus* (group VII). Two viruses were selected from group IV on the assumption that both were likely to share a larger number of attributes with the keyword compared to the viruses from other groups, which could thus contribute to evaluate the performance of the OEM more realistically. It is important to notice that the selection of these viruses was done randomly and without considering any potential semantic or class-affinity among the units involved in the experiment.

Finally, we manually selected a list of attributes or "descriptors" from the ExPASy Resource Portal for the two subordinate viruses from group IV in the Baltimore classification (Gasteiger et al., 2003)[2]. The portal contains detailed information of

[2] https://www.expasy.org/

every known virus, including its epidemiology, common interactions or natural hosts. Due to the complexity of the database, and to restrict the number of features for the experiment, we selected only the attributes in the “epidemiology” section, as shown in Figure 1[3].



TRANSMISSION
Zoonosis, arthropod bite, mainly mosquitoes

VACCINE
Japanese encephalitis virus, Tick-borne encephalitis virus, West Nile virus, Yellow fever virus.

ANTIVIRAL DRUGS
Ribavirin
YFV, WNV, Experimental: Favipiravir

Figure 1. *“Epidemiology” section in the ExPASy database.*

Consequently, the inventory of attributes for a virus such as *flavivirus* was as follows:

(1) arthropod, bite, borne, encephalitis, Favipiravir, fever, Japanese, mosquito, Nile, Ribavirin, Tick, virus, West, WNV, Yellow, YFV, Zoonosis

Although the selection was carried out entirely by manual means, the text was processed with the Data Mining Encountered (DAMIEN) tool (Periñán-Pascual, this volume)[4]. DAMIEN is a workbench included in the FunGramKB Suite (i.e. FunGramKB management interface) which offers users various options for corpus processing (e.g. tokenization, lemmatization, splitting, etc.), statistical processing (e.g. descriptive analysis, association measures, correlation, etc.), and data mining tools, including classification algorithms (e.g. decision trees, naïve Bayes, etc.) as well as clustering algorithms (e.g. K-means).

Future developments of the OEM, however, will require more advanced strategies for the extraction of a larger number of descriptors. In this regard, one avenue for research would consist in the extraction of semantic profiles from the batch processing of machine-readable sources such as the Wikipedia by using so-called

[3] This image was obtained from http://viralzone.expasy.org/all_by_species/24.html

[4] http://www.fungramkb.com/nlp.aspx

"start lists". In the case of virology, for example, a start list would allow to retrieve properties that are exclusively relevant to the specific domains of biology or medicine. As for their implementation, start lists can be made up on the basis of expert counselling or by importing terms from existing databanks such as the InterActive Terminology for Europe (IATE) term database, which offers datasets of concepts for most common branches of knowledge such as politics, commerce or transport[5].

4.2. Task 2: Corpus compilation

The second task involves the compilation of a corpus representative of the domain ontology. Whilst the lexicographic and/or encyclopedic sources provided the baseline for the retrieval of superordinate attributes, as seen in Section 2.1, the corpus will be used to obtain the finite set of attributes of the input keyword to be classified (i.e. *dengue*).

The application of the OEM to expert areas of knowledge necessarily requires a specialized corpus that must be the result of a balanced selection of texts. For the experiment, we used the Corpus of Contemporary American English (COCA), which contains more than 520 million words, and more specifically, we retrieved a corpus sample from the academic component, which contains approximately 103 million words of written text from scholarly journals (Davies, 2008-). We then selected 500 occurrences of the word *dengue* in this specialized section of the corpus and carried out pre-processing tasks with DAMIEN to reduce irrelevant n-grams in the sample. In this regard, the OEM relies on lexical filters for the removal of both functional words (e.g. pronouns, conjunctions, etc.) and non-alphabetical characters (e.g. numbers and symbols) as well as on common-language stopword lists (e.g. *individual*, *important*, etc.). A further filter in the case of expert corpora consists in the application of closed lists of high-frequency n-grams related to academic language (*analysis*, *experiment*, etc.) and metalanguage (*abstract*, *references*, etc.), as this type of vocabulary is recurrent in the scientific discourse. Finally, the sentences were tokenized as a list of unigrams and the Normalized Pointwise Mutual Information score (Bouma, 2009) was calculated to find the collocates of *dengue* (see Figure 2).

$$NPMI\ (x;y) = \frac{\ln \frac{p(x,y)}{p(x)p(y)}}{-\ln \mathrm{p}(x,y)}$$

Figure 2. *Normalized Pointwise Mutual Information formula.*

[5] http://iate.europa.eu/

The NPMI is an association metric which is used in linguistics to measure the semantic attraction between two variables, for example two lexical units. In other words, it measures whether the frequency of co-occurrence of both variables is statistically significant or is due to chance alone. The NPMI is mathematically expressed as a score within the range from 0 to 1, with 0 indicating absolute independence of both expressions with respect to each other and 1 indicating very significant co-occurrence.

4.3. Task 3: Data classification

The last step of the OEM involves the automatic classification of the input keyword by assigning it to one parent concept in the ontology. The classification requires two data structures: the training set and the test set. In the present experiment, the former consisted of all the features of the superordinates in the ontology, which were obtained from the ExPASy database, while the test set contained the features of the input keyword retrieved from the corpus. In fairly basic mathematical terms, the OEM relies on a function assigning a statistical weight to the unit to be classified, so that the higher the score the more relevant the unit is in defining a superordinate concept in the ontology. Before the function could be applied to our experiment, however, it was necessary to carry out two preparatory tasks. First, we used SQL (Structured Query Language) commands to find coincidences between the attributes of *dengue* and the attributes of each subordinate virus. As a result, all the attributes that were not found in the *dengue* vector were excluded from each of the individual virus vectors. Second, we manually established nine intervals or “ranges” within the *dengue* vector, so that the first range contained the collocates with the highest NPMI value, whilst the ninth contained the range of words that were least semantically attracted to *dengue*. It is noteworthy to mention that this segmentation of significance levels was carried out in a principled way by taking as the eligibility criterion any observable variations (i.e. significant increase or decrease) in the NPMI scores. Finally, we assigned a specific statistical weight to each of the nine ranges, from 0.9 for the topmost collocates in range 1 to 0.1 for the weakest collocates in range 9 (Table 2):

Range	**NPMI interval**	**Weight**	**Example attributes**
1	0.80 > x > 0.63	0.9	*fever, malaria, virus*
2	0.62 > x > 0.60	0.8	*epidemic, incidence, yellow*
3	0.59 > x > 0.57	0.7	*infection, hemorrhagic, serotype*

4	0.56 > x > 0.54	0.6	*confirm, encephalitis, endemic*
5	0.53 > x > 0.51	0.5	*Asia, temperature, west*
6	0.50 > x > 0.48	0.4	*cholera, Nile, serum*
7	0.47 > x > 0.45	0.3	*antibody, august, dysentery*
8	0.44 > x > 0.42	0.2	*geographic, diagnosis, evacuation*
9	0.41 > x > 0.30	0.1	*altitude, Caribbean, larva*

Table 2. *Weight variation of the* dengue *attributes.*

Section 5 presents the main results obtained after the OEM was applied to the corpus sample and the attribute vector of superordinates.

5. Results

The processing of both the corpus of the input keyword and the superordinates produced nine tables of attributes. To illustrate, Table 3 shows the list of collocates of *dengue* ranked by the NPMI score, so that the strongest collocates appear in the first positions:

Word	**NPMI**
fever	0.8044
case	0.7033
virus	0.7003
disease	0.6581
malaria	0.6564
mosquito	0.6559
vector	0.6355
outbreak	0.6299
epidemic	0.6071
incidence	0.6060

yellow	0.6060
patient	0.6036
cause	0.5973
report	0.5967

Table 3*. Topmost collocates of the* dengue *keyword.*

As regard the viruses in the ontology, two showed the highest weight scores in relation to the *dengue* attribute vector: *flavivirus* (Table 4) and *deltaretrovirus* (Table 5). The other viruses presented medium to small weight variations and therefore less semantic affinity with *dengue*.

Attribute	**NPMI**	**Range of *dengue***	**Weight**
fever	0.8044	1	0.9
virus	0.7003	1	0.9
mosquito	0.6559	1	0.9
yellow	0.6060	1	0.9
encephalitis	0.5444	4	0.6
west	0.5170	5	0.5
Nile	0.5090	6	0.4
Japanese	0.4947	6	0.4
bite	0.4518	7	0.3
Total			5.8

Table 4. *Statistical weights of the attributes of* flavivirus.

Attribute	**NPMI**	**Range of *dengue***	**Weight**
virus	0.7003	1	0.9
infection	0.5908	3	0.7
transmission	0.5607	4	0.6
result	0.5533	4	0.6

tropical	0.5444	4	0.6
[et] al	0.5090	6	0.4
fatal	0.4947	6	0.4
Total			4.2

Table 5. *Statistical weights of the attributes of* deltaretrovirus.

In order to compare the semantic relevance of each virus against the *dengue* virus, the results were normalized using a z-score (Table 6).

Virus	Weight	Z-score
aviadenovirus	2.4	0.84
circovirus	0.4	-1.15
deltaretrovirus	4.2	2.64
flavivirus	5.8	4.24
henipavirus	2.8	1.24
hepacivirus	0.6	-0.95
orthohepadnavirus	3.2	1.64
phytoreovirus	2.9	1.34

Table 6. *Weight and Z-score of the sample viruses from the Baltimore's taxonomy.*

As shown in Table 6, *flavivirus* scored the highest among the candidate host viruses, with a total z-score of 4.24. The interpretation of these results and its implications for the classification task set out in the initial hypothesis will be discussed in the next section.

6. Discussion

The results in Section 5 indicate a strong association between *dengue* and the genus *flavivirus* both in the number of shared features and the normalized statistical similarity between both. It can be thus concluded that the former belongs to group IV in the Baltimore classification, as predicted from the gold standard in Table 1. Qualitatively, these results allow validating the initial hypothesis that the OEM can be successfully applied to an input sample in allocating new concepts in a reference taxonomy. The OEM also showed a good performance insofar as it combined correctly the features of the two group-IV viruses that were purposely selected during the ontology selection (cf. Section 2.1) assuming that this would add extra difficulty to the process of automatic classification.

At a more general level, the results of the experiment show that the morphosyntactic distribution of keywords and phrases play a fundamental role in the semantic definition of both, which should be given greater prominence in future computational systems for natural language processing. This also seems to follow from the fact that the co-occurrences in our corpus sample contained core defining attributes which decidedly contributed to the correct classification of *dengue*. Consequently, the proposed OEM as well as other ontology-expansion systems must continue to explore the lexical neighborhood of keywords, even beyond the scope of single sentences, as a fundamental ontological principle. Also, lexical co-occurrence together with greater textual data will allow higher dimensional attribute spaces and therefore more accurate classifications.

A final question remains whether the OEM may be used in the population of other fields of knowledge and the enrichment of common-sense taxonomies. In this regard FunGramKB is a common-sense knowledge base that distinguishes three conceptual levels (i.e. metaconcepts, basic concepts and terminal concepts) which could clearly benefit from the described OEM (Periñán-Pascual & Mairal, 2009, 2010; Periñán-Pascual, 2013; Periñán-Pascual & Arcas, 2010; Mairal & Ruiz de Mendoza, 2006; Mairal & Periñán-Pascual, 2016). FunGramKB is structured in three modules: the Ontology, the Cognicon and the Onomasticon. The Ontology is in turn subdivided into two modules: a general-purpose module (Core) and the domain-specific modules (Satellites). The Ontology stores semantic knowledge, structured as a hierarchy of concepts whose properties are expressed in terms of meaning postulates. The Cognicon stores procedural knowledge by means of scripts (sequences of stereotypical actions in chronological course). Finally, the Onomasticon stores information about instances of entities and events. At present FunGramKB contains a hierarchy of about two thousand concepts, each of which is connected to a computationally-tractable semantic definition, as well as to a lexical inventory for languages such as English, Spanish, Italian or German. The knowledge base, however, should not be conceived as a closed system but it should be flexible enough to incorporate any new concept that may be created by speakers of different languages. Furthermore, it should be accessible to any lexico-semantic refinements which may be relevant to the units that are already part of the ontology. In the case of neologisms, for example, the OEM could be implemented as part of the FunGramKB Suite allowing the core hierarchy to be expanded with new terminal concepts as well as allowing near synonyms to be introduced into the lexical component. The examples of lexico-conceptual expansion are in fact almost limitless. For instance, the OEM should be able to allocate recently-coined concepts such as *crowdfunding*, *algocracy* or *verbicaine* under the relevant concepts PAY, GOVERNMENT or SAY in FunGramKB, respectively[6].

7. Conclusion

The main contribution of this paper has been to present a proof-of-concept method for expanding ontologies from expert domains of knowledge. The preliminary findings show that, even with a limited number of data, the method can be successfully applied for the automatic classification of specialized units into a reference ontology. Its main advantage is that it draws on corpus and terminological data which can be accessed from academic and encyclopedic sources. The "ontology enrichment method" (OEM), as we have termed it, opens new avenues for research in the processing of natural language, which is massively being produced online every day.

The paper has presented the results of a small-scale experiment which tested the capacity of the OEM to classify an "input keyword" automatically under its corresponding superordinate in a virus ontology. For this purpose, a set of attributes was first collected of each superordinate from a small ontology using a specialized database. Likewise, a corpus was compiled to obtain the attributes of the input virus *dengue* which was then matched against the superordinate attributes so as to detect significant semantic correlations. The matching between both sets was carried out heuristically by assigning statistical weights to various ranges in the attribute vector of *dengue*, so that an attribute was more significant as it ranked higher in the scale of attributes of the input keyword. The results showed that *dengue* had a higher similarity index with the class *flavivirus*, both as regards the number of shared features and as regards the position that these occupied among the strongest collocates of *dengue*.

There are three main issues that require further research. First, the experiment included a very reduced number of data: only one item to be classified and only eight concepts from the ontology. In order to check the general validity of the method therefore a wider number of lexical units must be considered, including

[6] According to the Word Spy neologisms dictionary (http://www.wordspy.com): a) the noun *crowdfunding* means "getting a large group of people to finance a project by using a website or other online tool to solicit funds"; b) the verb *verbicaine* means "soothing words used to calm or distract a patient who is awake during a surgical procedure"; and c) the noun *algocracy* means "rule or government by algorithm".

unigrams as well as bigrams and trigrams. Second, further research is needed to apply the OEM to non-specialized language so as to export the model to the enrichment of general-domain ontologies. Common language constitutes a useful benchmark to test the robustness of the method, since it involves computationally-complex phenomena such as ambiguity and polysemy, which may increase the number of false positives during the classification process and may thus challenge the methodological criteria adopted in this paper. Finally, some improvements have been mentioned in this paper for achieving higher automatization rates of attribute selection that are essential to process large quantities of natural language. The experiment – as it stands – constitutes a preliminary version of the OEM and therefore a great deal of manual work is involved prior to the automatic classification phase. Future developments should therefore include the application of complex unsupervised algorithms to the training set which can manage datasets with thousands of attributes.

About the authors

Pedro Ureña Gómez-Moreno is Assistant professor at the Department of Didactics of Language and Literature at the University of Granada (Spain), where he develops most of his teaching and research activity. His teaching focuses on Natural Language Processing, Corpus Linguistics and English as a Second Language, both at the University of Granada and the UNED. His main areas of research are Morphosyntax and Lexicology within the frameworks of Corpus Linguistics and Natural Language Processing, with a special interest in Terminology and Knowledge Engineering applied to the development of FunGramKB Knowledge Base. A second line of research concerns the application of new technologies to language teaching and the development of virtual courses. He has authored and co-authored a number of refereed book chapters in Mouton de Gruyter and John Benjamins, as well as several articles in national and international journals, including The International Journal of Corpus Linguistics, Onomázein or The LSP Journal.

Eva M. Mestre-Mestre works as associate professor at Universitat Politècnica de València. Since her Ph.D. thesis on the pragmatic implications of errors in English as a second language, her research has focused on Pragmatics, English learning in higher education, and corpus management, including computational linguistics, resulting in publications indexed in nationally and internationally prestigious journals, such as RESLA, or the Yearbook of Pragmatics. Apart from several book chapters, she has co-edited Understanding Meaning and Knowledge

Representation for Cambridge Scholars Press. She was a visitor researcher in several European and American universities. She is currently the director of the panel on pragmatics in the Spanish Society for Applied Linguistics, and director of the panel on ESP in the Spanish Society for Corpus Linguistics.

Acknowledgements

Financial support for this research has been provided by the DGI, Spanish Ministry of Education and Science, grant FFI2011-29798-C0201.

Article history

Paper received: 20 February 2017
Paper received in revised form and accepted for publication: 27 April 2017

References

Alfonseca, E. & Manandhar, S. (2002). Extending a lexical ontology by a combination of distributional semantics signatures. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management* (pp. 1–7). Berlin: Springer

Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriological Review, 35*(3), 235-241.

Bendaoud, R., Toussaint, Y. & Napoli, A. (2008). PACTOLE: A methodology and a system for semi-automatically enriching an ontology from a collection of texts. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5113 LNAI, 203–216.

Biemann, C. (2005). Ontology learning from text: A survey of methods. *LDV-Forum, 20*(2), 75–93.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In W. Brinkman, J. Broekens & D. Heylen (Eds.), In *Proceedings of the Biennial GSCL Conference* (pp. 31-40). Potsdam.

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing* (pp. 152-155).

Ciaramita, M., Gangemi, A., Ratsch, E., Saric, J. & Rojas, I. (2005). Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 659–664). Professional Book Center.

Davies, M. (2008-). The corpus of contemporary American English (COCA): 520 million words, 1990-present. <http://corpus.byu.edu/coca/> [24/03/2017].

Dorr, B. & Jones, D. (1996). Acquisition of semantic lexicons: Using word sense disambiguation to improve precision. In *Proceedings of the SIGLEX Workshop on Breadth and Depth of Semantic Lexicons*, (pp. 42–50).

Faatz, A. & Steinmetz, R. (2002). Ontology enrichment with texts from the WWW. In the Semantic Web Mining Conference, WS02.

Gacitua, R., Sawyer, P. & Rayson, P. (2008). A flexible framework to experiment with ontology learning techniques. *Knowledge-Based Systems, 21*(3), 192–199.

Gamallo, P., Gonzalez, M., Agustini, A., Lopes, G. & Delima, V. (2002). Mapping syntactic dependencies onto semantic relations. In *Proceedings of the ECAI Workshop on Machine Learning and Natural Language Processing for Ontology* (pp. 15-22).

Gasteiger E., Gattiker A., Hoogland C., Ivanyi I., Appel R. D. & Bairoch A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research, 31*(13), 3784–3788.

Gómez-Pérez, A. & Manzano-Macho, D. (2004). An overview of methods and tools for ontology learning from texts. *The Knowledge Engineering Review, 19*(3), 187–212.

Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), 199–220.

Gupta, K. M., Aha, D., Marsh, E. & Maney, T. (2002). An architecture for engineering sublanguage WordNets. In *Proceedings of the First International Conference on Global WordNet* (pp. 207–215). Central Institute of Indian Languages, Mysore.

Haase, P. & Stojanovic, L. (2005). Consistent Evolution of OWL Ontologies. In A. Gómez-Pérez & J. Euzenat (Eds.), *ESWC 2005. LNCS*, vol. 3532, (pp. 182–197). Heidelberg: Springer.

Hahn, U. & Schnattinger, K. (1998). Towards text knowledge engineering. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence American Association for Artificial Intelligence* (pp. 524–531).

Hippisley, A., Cheng, D. & Ahmad, K. (2005). The head-modifier principle and multilingual term extraction. *Natural Language Engineering, 11*(2), 129–157.

Hotho, A. Madche, A. & Staab, S. (2001). Ontology-based text clustering. In *Proceedings of the IJCAI-2001 Workshop Text Learning: Beyond Supervision* (pp. 48–54). Seattle, USA.

Hwang, C. (1999). Incompletely and imprecisely speaking: Using dynamic ontologies for representing and retrieving information. In *Proceedings of the 6th International Workshop on Knowledge Representation meets Databases*.

Ide, N. & Véronis, J. (1998). Word sense disambiguation: The state of the art. *Computational*

Linguistics, 24(1), 1–40.

Khan, L. & Luo, F. (2002). Ontology construction for information selection. In *Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence* (pp. 122–127). Crystal City, Virginia.

Klein, D. & Manning, C. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics* (pp. 423-430).

Lee, C. S., Kao, Y. F., Kuo, Y. H. & Wang, M. H. (2007). Automated ontology construction for unstructured text documents. *Data and Knowledge Engineering, 60*(3), 547–566.

Lee, J., Kim, J. & Park, J. (2006). Automatic extension of gene ontology with flexible identification of candidate terms. *Bioinformatics, 22*(6), 665–670.

Lima, R., Oliveira, H., Freitas, F. & Espinasse, B. (2014). Ontology population from the web: An inductive logic programming-based approach. ITNG 2014. In *Proceedings of the 11th International Conference on Information Technology: New Generations* (pp. 473–478).

Lin, D. (1994). Principar: An efficient, broad-coverage, principle-based parser. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)* (pp.482-488). Kyoto, Japan.

Lisi, F. A. (2005). Principles of inductive reasoning on the semantic web: A framework for learning in AL-Log. In F. Fages, & S. Soliman (Eds.), *PPSWR 2005. LNCS*, vol. 3703 (pp. 118–132). Heidelberg: Springer.

Liu, W., Weichselbraun, A. & Chang, E. (2005). Semi-automatic ontology extension using spreading activation. *Journal of Universal Knowledge Management, 0*(1), 50–58.

MacNamara, J. (1982). Names for things: A study of human learning. Cambridge, MA: MIT Press.

Missikoff, M., Navigli, R. & Velardi, P. (2002). Integrated approach to Web ontology learning and engineering. *IEEE Computer, 35*(11), 60–63.

Mitchell T. (1997). Machine Learning. New York: McGraw-Hill.

Navigli, R. & Velardi, P. (2002). Automatic adaptation of WordNet to domains. In *Proceedings of 3rd International Conference on Language Resources and Evaluation* (pp. 1023-1027).

Navigli, R. & Velardi, P. (2004). Learning domain ontologies from document warehouses and dedicated web sites. *Computational Linguistics, 30*(2), 151–179.

Navigli, R., Velardi, P., Cucchiarelli, A. & Neri, F. (2004). Quantitative and qualitative evaluation of the OntoLearn ontology learning system. In *Proceedings of the 20th International Conference on Computational Linguistics.*

O'Hara, T., Mahesh, K. & Nirenburg, S. (1998). Lexical acquisition with WordNet and the Mikrokosmos Ontology. In *Proceedings of the ACL Workshop on the Use of WordNet in NLP* (pp. 94–101).

Pedersen, T., Patwardhan, S. & Michelizzi, J. (2004). WordNet:similarity: Measuring the relatedness of concepts. In *Proceedings of the Demonstration Papers at the Conference of* *the North American Chapter of the Association for Computational and Linguistics: Human Language Technologies (HLT-NAACL)*.

Periñán-Pascual, C. (2013). Towards a model of constructional meaning for natural language understanding. In B. Nolan & E. Diedrichsen (eds.) *Linking constructions into Functional Linguistics: The role of constructions in grammar* (pp. 205–230). Amsterdam/Philadelphia: John Benjamins.

Periñán-Pascual, C. (2017). Bridging the gap within text-data analytics: A computer environment for data analysis in linguistic research. *Revista de Lenguas para Fines Específicos*, *23*(2), 111-132.

Periñán-Pascual, C. & Arcas Túnez, F. (2010). The architecture of FunGramKB. *7th International Conference on Language Resources and Evaluation*, Valeta (Malta). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA)* (pp. 2667–2674).

Periñán-Pascual, C. & Mairal Usón, R. (2009). Bringing Role and Reference Grammar to natural language understanding. *Procesamiento del Lenguaje Natural 43*, 265–273.

Periñán-Pascual, C. & Mairal Usón, R. (2010). Enhancing UniArab with FunGramKB. *Procesamiento del Lenguaje Natural, 44*, 19–26.

Poon, H. & Domingos, P. (2010). Unsupervised ontology induction from text. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 296–305).

Princeton University (2010) About WordNet. *WordNet*. Princeton University <http://wordnet.princeton.edu> [24/03/2017].

Roussinov, D. & Zhao, J. L. (2003). Automatic discovery of similarity relationships through web mining. *Decision Support Systems, 35*(1), 149–166.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing Management, 24*(5), 513–523.

Salton, G., Wong, A. & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620.

Shamsfard, M. & Barforoush, A. (2004). Learning ontologies from natural language texts. *International Journal of Human-Computer Studies, 60*(1), 17–63.

Studer, R., Benjamins, V. R. & Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering, 25*(1-2), 161–197.

Udrea, O. & Getoor, L. (2007). Combining statistical and logical inference for ontology alignment. In *Proceedings of the Workshop on Semantic Web for Colaborative Knnowledge Acquisition, IJCAI* (pp. 51–58). Hyderabad, India.

Widdows, D., Dorow, B. & Chan, Ch. (2002). Using parallel corpora to enrich multilingual lexical resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation* (pp. 240–245). Las Palmas, Spain.

Wong, W., Liu, W. & Bennamoun, M. (2012). Ontology learning from text. *ACM Computing*

Surveys, 44(4), 1–36.

Word Spy. Logophilia Limited. < https://www.wordspy.com/> [24/03/2017].