Applied Ontology 7 (2012) 429–448
DOI 10.3233/AO-2012-0114
IOS Press

Corpus-based terminological evaluation of ontologies

Marco Rospocher *, Sara Tonelli, Luciano Serafini and Emanuele Pianta
Fondazione Bruno Kessler-irst, Via Sommarive 18 Povo, I-38123, Trento, Italy
E-mails: {rospocher, satonelli, serafini, pianta}@fbk.eu

Abstract. We present a novel system for corpus-based terminological evaluation of ontologies. Starting from the assumption that a domain of interest can be represented through a corpus of text documents, we first extract a list of domain-specific key-concepts from the corpus, rank them by relevance, and then apply various evaluation metrics to assess the terminological coverage of a domain ontology with respect to the list of key-concepts.

Among the advantages of the proposed approach, we remark that the framework is highly automatizable, requiring little human intervention. The evaluation framework is made available online through a collaborative wiki-based system, which can be accessed by different users, from domain experts to knowledge engineers.

We performed a comprehensive experimental analysis of our approach, showing that the proposed ontology metrics allow for assessing the terminological coverage of an ontology with respect to a given domain, and that our framework can be effectively applied to many evaluation-related scenarios.

Keywords: Corpus based ontology evaluation, terminological ontology evaluation, key-concept extraction, ontology building environment

1. Introduction

An ontology is a formal conceptualization of some domain of interest. Ontologies are increasingly used for organizing information in several application fields, including among others the Semantic Web, knowledge representation and management, biomedical informatics, software engineering, and enterprise management. Several methodologies and tools are available to support building and developing ontologies (see Gómez-Pérez et al., 2004, for a detailed overview of the Ontology Engineering field).

As any engineering artifact, an ontology needs to undergo some exhaustive evaluation, for example to understand whether it adequately describes a given domain of interest, or to check whether it is formally correct. *Ontology evaluation* is the task of investigating the quality of an ontology. The investigation can concern different levels (as summarized in the survey provided in Brank et al., 2005), such as the *terminological level*[1] (*Does the ontology represent the relevant terms of the domain of interest?*), the *syntactic level* (*Does the ontology match the syntactic requirements of the formal language adopted?*), the *hierarchical or taxonomical level* (*Does the ontology structurally fit the domain of interest?*) and the *semantic level* (*Does the underlying semantic model in the ontology correctly represent the domain of interest?*).

*Corresponding author. E-mail: rospocher@fbk.eu.

[1]We prefer to adopt the term *terminological* in place of *lexical/vocabulary* used in Brank et al. (2005).

The contribution presented in this paper concerns the terminological level, since it aims at assessing whether an ontology *adequately covers* the domain of interest, i.e. whether the concepts used in the ontology comprehensively represent the relevant terms of a domain.

More specifically, we present a *framework for the corpus-based terminological evaluation of ontologies*, where an ontology is terminologically evaluated against a text corpus representative of the domain of interest. The approach is based on the extraction of a list of relevant concepts (aka *key-concepts*) from a domain corpus, ranked according to their relevance, and a matching-based comparison (aka *matching*) between the concepts formalized in the ontology and the extracted key-concepts. To obtain a more accurate result, the matching relies on synonymy information available in WordNet (Fellbaum, 1998). Based on the resulting matching, several evaluation metrics are defined to assess whether the given ontology adequately covers the terminology of the domain described by the text corpus.

The terminological evaluation of ontologies has been widely studied in the literature (we refer the reader to Section 2 for an overview of available proposals, and a comprehensive comparison between them and our approach); nevertheless our contribution is novel under several aspects, and presents many advantages over other state of the art proposals:

High level of automation: human intervention is limited to the selection of the reference corpus and, possibly, to the tuning of the terminology extraction module. There is not need for a manually built gold standard;

On-line evaluation environment: a collaborative system fully implementing the proposed evaluation framework has been developed and made publicly available.[2] Therefore, users can exploit our framework to terminologically evaluate any OWL ontology against any text corpus (several popular digital text file formats are supported);

Domain independence: domain-specific language resources are not required;

Weighted coverage assessment: thanks to the relevance-based ordering of the key-concepts extracted from the corpus, the methodology can be applied to assess whether the most important concepts in the domain (as opposed to the marginal ones) are covered by the ontology.

We perform a comprehensive experimental analysis of our approach, showing that the evaluation metrics proposed appropriately capture the terminological adequacy of an ontology with respect to a domain. Such metrics can be employed also to effectively and efficiently rank candidate ontologies according to how they terminologically cover a given domain, or to understand which are domain-wise the most relevant concepts formalized in an ontology.

We remark that the contribution here presented allows to evaluate the *terminological* level of the ontology, i.e. whether the terms used as concepts in the ontology are the relevant terms of a domain of interest, while it does not deal with the evaluation of the *semantic* level of the ontology, that is whether the axiomatization of the domain encoded in the ontology (i.e. the OWL axioms characterizing the concepts and properties in the ontology) is correct and complete.

The paper is structured as follows. In Section 2 we present a comprehensive overview of available proposals for ontology evaluation at terminological level. In Section 3 we describe our corpus-based ontology evaluation framework together with the ontology metrics we propose to adopt, while in Section 4 we describe the collaborative system we developed to implement the proposed approach. Furthermore, in Section 5 we detail some application scenarios in which our corpus-based evaluation framework can be effectively applied. In Section 6 we report the detailed experimental analysis that we performed to

[2]To the best of our knowledge, this is the first collaborative system of this kind made publicly available.

evaluate our framework, while in Section 7 some limitations of our approach are discussed. Finally, we draw some conclusions in Section 8 and present future research directions we plan to undertake.

2. Related work

Ontology evaluation can be based on different approaches. One of them is the *manual* revision by experts, which however has several drawbacks, being time-consuming and sensitive to the subjective nature of human interpretation and judgement. Some tools have been developed to support the user in manual ontology revision by assigning weights and values to the dimensions characterizing an ontology, for example the OntoMetric Tool (Lozano-Tello & Gómez-Pérez, 2004) and the COAT tool (Bolotnikova et al., 2011). The latter is focused on the evaluation of the cognitive ergonomicity of ontologies, i.e. on aspects concerning the human speed of perception and the cognitive soundness.

As for automatic evaluation, some attempts have been made to define appropriate standards and requirements. A well-studied approach is the evaluation of the ontology against a reference ontology, aka *gold standard*. Many metrics have been proposed to compare ontologies both at lexical and conceptual level: Maedche and Staab (2002) measure both the lexical overlap between concept names and the taxonomic structure of two ontologies in an empirical study on the tourism domain, while Dellschaft and Staab (2006) suggest a number of criteria for evaluation as well as several measures of similarity. Note that although the comparison between the evaluated ontology and the gold standard can be easily automatized, the building of the gold standard is still manual.

A further approach for ontology evaluation is *application-based* in that it measures the quality of an ontology based on the improvement achieved by an application that is built upon it. Porzel and Malaka (2004), for example, evaluate the accuracy of an ontology by integrating it in a system for relation tagging.

In this work, we propose a methodology to evaluate an ontology based on a *domain corpus*. Few works go in this direction, and no evaluation system has been made available so far. In Brewster et al. (2004) the authors present a data-driven methodology for evaluating an ontology by comparing it with a corpus representing the domain area. This approach is the most similar to ours, since it is based upon the same principle, i.e. a domain corpus can be used as a starting point to evaluate the terminological adequacy of an ontology representing the knowledge of the same domain. The authors present a first evaluation methodology based on a vector space representation of the terms shared by an ontology and a corpus. However, the corpus is built by collecting 41 arbitrary texts from the Internet concerning arts and artists, therefore it cannot be seen as a *reference* corpus for a domain. Besides, none of the five ontologies compared to the corpus has been independently evaluated, so no real evidence of the efficacy of this evaluation approach is given. The authors also present a more sophisticated methodology, proposing to measure the "fit" between an ontology and the corpus as the conditional probability of the ontology given a corpus. Although the approach seems very interesting, neither related experiments nor evaluation are reported.

Jones and Alani (2006) present a methodology inspired by Brewster et al. (2004), but select the corpus based on a Google query extended with WordNet terms. Tf-Idf (Term frequency/Inverse document frequency) is then applied to the corpus in order to extract the top 50 potential concept labels to match against the ontology. The authors show that their approach can be applied to rank 10 candidate ontologies according to the corpus domain, with high correlation with human judgement. However, their evaluation is focused only on the ranking and no attempts are made to find a relevance score that represents in

absolute terms the quality of the ontology with respect to the domain. Besides, both the corpus creation and the term extraction are quite simplistic and may require some further refinement.

More recently, Yao et al. (2011) present a methodology to benchmark an ontology against a reference corpus by first mapping concepts and relations to the corpus using NLP (Natural Language Processing) tools, and then estimating concept- and relation-specific frequency parameters to compute several similarity metrics between the ontology and the corpus. The authors rank five medical ontologies with respect to a medical corpus by taking into account precision and recall as well as the theoretical coverage and parsimony of ontology's concepts. The metrics rely on the *complete ontology* created by incorporating all concepts and relations found in the reference corpus, that represents some kind of gold standard. However, the process to create this complete ontology applies only to the medical domain, since it is based on the UMLP MetaMap. Our approach, instead, relies on a general purpose methodology, and the available system is able to deliver an evaluation for any ontology and domain corpus, given that they are in a suitable format.

Cui (2010) compares coverage, semantic consistency, and agreement of four plant character ontologies by checking them against domain literature. However, the approach has been developed for the biodiversity domain, and could not be applied to other domains, especially the semantic annotation algorithm used to extract character states.

3. The approach

In our evaluation framework, we assume that the knowledge domain that should be encoded in the ontology is represented through a domain corpus, and that the evaluation should output some measures that express the coverage and the adequacy of the ontology with respect to such domain. This is similar to the scenario presented by Brewster et al. (2004) and Cui (2010). For example, the corpus could consist of a document describing a certain knowledge field, or a collection of articles concerning a specific topic. Note that our approach works both with a corpus containing multiple documents and one formed by only a single (possibly long) text.

Given a corpus, the evaluation process is based on three steps, all performed in a pipeline without the need of human intervention:

(1) *Key-concept extraction*: Extraction of a ranked list of key-concepts from the corpus. Some manual tuning of the extraction algorithm is possible but not necessary;
(2) *Enrichment with external resources*: Enrichment of the ranked list with additional information (synonyms) from external resources (e.g. WordNet);
(3) *Matching & evaluation*: Alignment between the ontology and the enriched ranked list of key-concepts, and computation of some ontology metrics based on these alignments.

A graphical representation of the workflow is displayed in Fig. 1. The single steps are detailed in the following subsections.

3.1. Key-concept extraction

The first step aims at acquiring the terminology in the specific domain of interest, which is often seen as a useful starting point for supporting the creation of a domain ontology (see for example Liddle et al., 2003; Navigli and Velardi, 2004; Lee et al., 2005; Wong et al., 2007; for an overview, see Buitelaar et al., 2005). In fact, domain-specific terminology usually shows low ambiguity and high specificity,

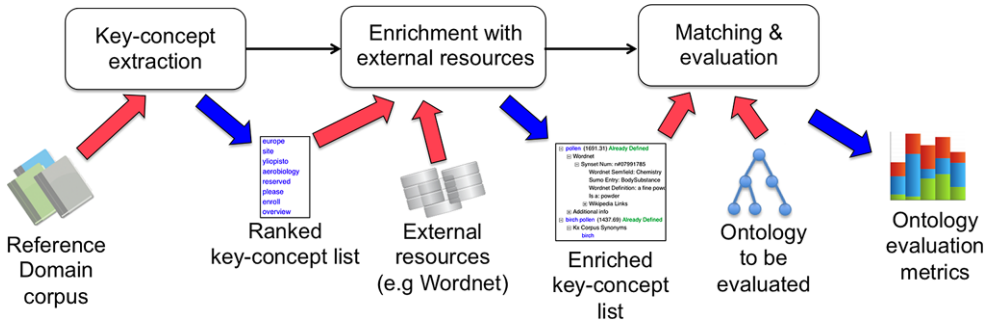

Fig. 1. Ontology evaluation workflow. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-2012-0114.)

therefore it can provide a concise and unambiguous representation of a knowledge domain. In this first processing step, however, we do not just extract the domain-specific terminology. Indeed, the terms are ranked according to their relevance in the corpus. Note also that, as opposed to the works mentioned above, key-concepts are not extracted with the goal of building an ontology, but for *evaluating* it.

We rely on KX (Pianta & Tonelli, 2010), a system that extracts from a single text or a corpus an ordered list of the most relevant key-concepts by combining statistical information and lexical-based filtering. Key-concepts, as implemented in KX, can in principle be interpreted as an equivalence class on synonym terms, abstracting not only over morphological and syntactic variance, but also over semantic equivalence. However, in order to recognize that two distinct terms occurring in a corpus represent occurrences of the same concept, knowledge (possibly domain-specific) about semantic equivalence must be provided to KX. Specifically, manual synonym lists can be integrated in the extraction workflow. Since we aim at proposing a domain-independent evaluation framework, we will not fully exploit this potentiality of the KX system. Thus, in the vast majority of cases, key-concepts should be considered as equivalent to terms, with some ability to abstract over morphological variance.

KX does not need an annotated corpus for training, as opposed to existing supervised approaches (see Witten et al., 1999, Hulth, 2004, among others). This implies that the system can be straightforwardly used through the online interface after a manual parameter tuning. Another approach to the unsupervised extraction of relevant concepts from a corpus could be explored using topic models such as latent Dirichlet allocation (Blei et al., 2002). LDA represents documents as a mixture of topics, and a topic is a distribution over words in a vocabulary. The advantage of this model is that the most relevant words in a corpus are recognized and clustered into topics, while the clustering step is not performed by KX. However, KX guarantees a better control over the number and the length of the key-concepts, which is crucial to the adaptation to different corpora and domains. Furthermore, it ranks the list of key-concepts by relevance.

The process performed by KX and leading to the extraction of an ordered list of terms is based on the following subtasks:

N-gram Extraction: The list of all possible n-grams ($n > 1$) ordered by frequency is extracted from the corpus. The maximum length of n-grams can be set by the user. N-grams with frequency in the corpus lower than a threshold are filtered out.

Multi-word Extraction: The n-gram list is filtered with the aim of retaining only Multi Word Expressions (MWEs), i.e. combinations of words expressing a unitary concept, for example 'business process diagram' or 'message flow'. Filtering is carried out through two mechanisms. First we ex-

clude all n-grams containing at least one of the words contained in a black list, e.g. 'because', 'is', etc. Then we retain only n-grams matching one of a predefined list of lexical patterns (e.g. Adjective + Noun, Noun + Preposition + Noun). Word sequences are selected as they are, without performing lemmatization. What we get from this filtering is an ordered list of probable MWEs.

Multi-word Recognition: With the list produced in the previous subtask, we go back to the corpus and recognize the actual occurrence of MWEs in each text, possibly solving local ambiguities between nested or partially overlapping MWEs. After multi-word recognition, the units of our text are not tokens anymore, but lexical units, where a lexical unit can include one or more tokens.

Ranking of key-concept by frequency: We group lexical units into key-concepts by recognizing some morphological variants and synonyms based on a manually compiled synonym list. For each text we can now build a list of key-concepts ordered by frequency. A manual black-list can be applied to avoid that certain specific concepts are included in the key-concept list of any text.

Re-ranking of key-concepts by relevance: For each text, the initial list of key-concepts ordered by frequency is re-ranked by taking into consideration various parameters that can be set by the user such as first occurrence of key-concepts in the text and degree of specificity (more details are provided later on in this section).

Output of final key-concept list: The single key-concept lists extracted at document level are merged in a unique list with normalized relevance (from 0 to 1). Only the key-concepts with a relevance >0.01 are returned and considered representative of the domain.

Since KX has been first developed as a standalone system that can be employed both for key-concept extraction from a single text and for terminology extraction from a corpus, many configurable parameters have been implemented in order to allow users to tailor the output to the document characteristics and to their specific needs. For instance, if the tool is used to obtain a possibly small set of key terms expressing the most relevant content of a document, a user may want to keep in the final list only very specific key-concepts. To this purpose, a parameter can be activated to boost more informative (longer) key-concepts while filtering out the *nested* ones (Frantzi et al., 2000), i.e. those that appear within longer key-concepts. For example, this parameter would assign high relevance to 'business process modeling notation' while discarding 'modeling notation'. To the same purpose, another parameter allows users to boost longer key-concepts by multiplying the relevance by their length in tokens.

In the framework of ontology evaluation, different parameter combinations were tested with the aim of finding the optimal degree of key-concept specificity. A high preference for specific key-concepts may reduce to zero the relevance of shorter (nested) ones, i.e. filter them out from the final ranking. This may not be appropriate if the extracted key-concepts should be representative of a domain, where both generic concepts and their specifications can occur (for example 'pollutants', 'air pollutants' and 'water pollutants' may all refer to the pollution domain). In order to enable the user to easily set these parameters, we implemented four degrees of specificity for the extracted key-concepts, ranging from no preference for specific key-concepts to maximal preference. Based on our experiments, the best option for ontology evaluation was *no preference*.

3.2. Enrichment with external resources

In the second step, the list of key-concepts (terms) extracted by KX from the reference corpus is possibly enriched with the synonyms found in WordNet (Fellbaum, 1998). In particular, for each key-concept we check if it occurs in WordNet and, if many synsets are available for a given key-concept, we perform a disambiguation step, so as to uniquely match a key-concept with a single synset in WordNet

hierarchy. The goal of this step is to acquire additional synonyms for the given key-concepts that should be exploited in the matching phase (see Section 3.3). Specifically, all lemmas in the synset that has been associated with a key-concept are considered as synonyms. Note that the terms added in this step may not be contained in the reference corpus.

Word sense disambiguation (WSD) is performed using WordNet::SenseRelate::WordToSet library (Pedersen et al., 2005), which has been developed in order to disambiguate a word given the textual context in which it appears. Specifically, a word is assigned the synset that is most related to its neighboring words in the text. Since this approach was originally devised to disambiguate words in the sentences in which they appear, we adapted it to our specific application scenario: we build a context by considering the 10 top-ranked key-concepts returned by KX, and disambiguate all other key-concepts in the list based on them.[3] The intuition behind this is that the top-ranked concepts should be the most representative for the domain, and therefore they should build the most appropriate context to disambiguate other concepts from the same domain. We choose this approach also because the collaborative platform we present should be domain-independent and work on any unseen corpus and ontology. This lead us to discard supervised WSD approaches, which suffer from low performance when applied to unseen domains.

3.3. Matching & evaluation

After the identification of synonymous terms, the key-concept list is matched against the ontology. The matching is based on a *string-wise match*[4] between each ontology concept and the key-concepts extracted from the corpus, including the corresponding synonyms. Currently, our work takes into account only perfect matches, while we do not consider partial matches based on edit distance, as proposed e.g. by Maedche (2002). This aspect will be integrated in the evaluation system in the near future, after a careful calibration of the partial similarity metrics (i.e. of how partial matches should be quantified).

Once the matching between the ontology and the ranked list of key-concepts is performed, we can compute several evaluation metrics based on the matching obtained. In fact, since we aim at evaluating the adequacy of the ontology in covering terminologically the given domain, we consider the enriched key-concept list as the set of concepts (ranked by relevance) characterizing the whole domain. Given the enriched ranked list K of key-concepts extracted from the corpus, the set C_{onto} of concepts in the ontology to be evaluated, and the set $K_{correct}$ of *matching concepts*, i.e. those belonging both to the ontology and the enriched key-concept list, we implement standard *Precision*, *Recall* and $F1$ measures as follows:

$$\mathit{Precision} = \frac{|K_{correct}|}{|C_{onto}|}, \qquad \mathit{Recall} = \frac{|K_{correct}|}{|K|}, \qquad F1 = \frac{2 \times \mathit{Precision} \times \mathit{Recall}}{\mathit{Precision} + \mathit{Recall}}.$$

The above metrics have been widely used for ontology evaluation, although they have been mainly employed for comparing a given ontology against a gold standard one (see e.g. Dellschaft & Staab, 2006). In this case, instead, they are applied to compare an ontology against a list of automatically extracted key-concepts.

[3] The length of the context window corresponds to a trade-off between disambiguation time and accuracy, given that Pedersen et al. (2005) report the best disambiguation performance over nouns using a window of 20 words.

[4] Normalized with respect to empty spaces and characters case.

Since a normalized relevance value is assigned by KX to each key-concept, we also introduce a variant of *Recall* and $F1$ taking into account this information. The novel $Recall_w$ and $F1_w$ are computed as follows:

$$Recall_w = \frac{Rel_{K_{correct}}}{Rel_K}, \qquad F1_w = \frac{2 \times Precision \times Recall_w}{Precision + Recall_w},$$

where $Rel_{K_{correct}}$ is the sum of the relevance scores associated with the key-concepts in $K_{correct}$, and Rel_K is the sum of the relevance scores associated to all the key-concepts in K. We introduce this measure because we want to exploit one of the strengths of KX, i.e. the availability of a key-concept list ranked by relevance, which allows a user to know which concepts are most important in the domain. While standard *Precision* and *Recall* do not take into account the ranking, $Recall_w$ and $F1_w$ are higher if the matching concepts in the ontology are among the top-ranked ones, while they are lower if the matching concepts are at the bottom of the key-concept list. This will be empirically confirmed in our experiments (see Section 6.4).

Also Jones and Alani (2006) presented a measure called *Class Match Score* (CMS) that takes into account an ordered list of terms extracted from a reference corpus. However, their approach differs from ours in that their corpus is formed by the first 100 pages returned by Google after querying a term, and the key-concept list is limited to the top 50 concepts obtained from such corpus ranked by Tf-Idf. The authors do not take into account terms expressed by multi-words but only by single tokens, which limits the efficacy of ontology evaluation, given that domain-specific concepts in English are often multi-word expressions. Furthermore, they arbitrarily select the top 50 concepts, while our cutoff value for the key-concept list is more flexible because we rely on relevance.

4. The evaluation environment

In order to allow users to experiment with the approach described in Section 3, a running version of the evaluation system has been made available.[5] The system has been effectively used for ontology extension (Tonelli et al., 2011), while the evaluation module is presented for the first time in this paper.

Four main components are part of the framework:

MoKi : The evaluation process can be performed after accessing MoKi (Ghidini et al., 2010), a MediaWiki-based tool[6] for modeling ontological and procedural knowledge in an integrated manner.[7] The main idea behind MoKi is to associate a wiki page to each basic entity of the ontology, i.e. concepts, object and datatype properties, and individuals. Each basic entity is associated with a MoKi page, composed of an unstructured and a structured part. The unstructured part contains informal text, possibly enriched by formatting information, links to other MoKi pages or to external resources, uploaded images, and so on. In the current implementation, the content of this part is stored according to the standard MediaWiki markup format. The structured part, which is delimited by specific tags to separate it from the unstructured text, contains knowledge stored according to the

[5] http://moki.fbk.eu/moki/tryitout2.0/.

[6] See http://www.mediawiki.org.

[7] Though MoKi allows to model both ontological and procedural knowledge, here we will limit our description only to the features for building ontologies.

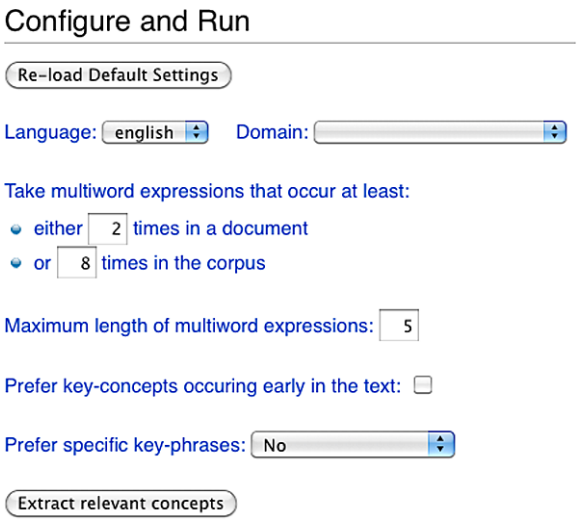

Fig. 2. KX parameters. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-2012-0114.)

modelling language adopted. In the current implementation, the structured part of a page contains a RDF/XML serialisation of a set of OWL statements formalising the element described in the page. MoKi implements a multi-mode access to the page content, to support easy usage both by domain experts (who own knowledge about a domain, sometimes implicitly, but usually lack the skill of making this knowledge explicit in a formal model) and knowledge engineers (who have the capability of encoding a piece of informal knowledge into a formal model, but usually have no or limited understanding of the domain to be modeled), thus facilitating them to play an equally central role in the modelling activities.

Before starting the evaluation, the user needs to upload in MoKi the ontology to be evaluated (in OWL format) as well as the domain corpus to be used (multiple files in popular file formats - plain text, Adobe PDF, Microsoft Office, Open Office – can be uploaded simultaneously).

KX: The key-concept extraction process is performed by KX, which can be activated through the MoKi interface after setting some available parameters (see Fig. 2).

First of all, the corpus language and the domain should be chosen. The possible parameters for the languages are 'English' and 'Italian', with English being the default value. The language choice is constrained by the languages supported by KX. We are currently extending the set of available languages to include French, Finnish and Swedish.

The domain options currently include 'Environment', 'Medical' and 'News'. This means that some large corpora for the given domain and language combinations have been pre-processed and that some relevant statistical information, for example key-concept IDF, have been made available in background. If this information is used, KX will boost the relevance of domain-specific key-concepts. When available, domain information is combined with information extracted from the target corpus. If domain information is not available, the terminology extraction process relies only on the target corpus. Note that in the experiments presented in Section 6 below, no domain information was used. Furthermore, the user can decide to select as key-concepts only the MWEs that occur at least n-times in each document belonging to the corpus or m-times in the whole corpus. The default values of 2 and 8 should be increased in case of long documents or large corpora. Also, the maximum length of MWEs can be set. A higher value means that longer (and more specific)

MWEs will be included in the term list. For ontology evaluation, this was set to five in order to include as much specific key-concepts as possible.

The parameter 'Prefer key-concepts occurring early in the text' assigns a higher relevance to key-concepts mentioned at the beginning of the text. Finally, the option 'Prefer specific key-phrases' offers five options, namely 'No', 'Weak', 'Medium', 'Strong' and 'Max' preference, which boost to an increasing degree the relevance of more specific concepts. After some tests, the best combination of the above parameters includes no preference for specific key-phrases and no boosted relevance for key-concepts occurring early in text. In this way, the final list of ranked key-concept still contains longer as well as nested key-concepts, for example 'process diagram' and 'business process diagram', which are both relevant to the domain and are likely to be in a taxonomic relation. It is interesting to note that these two parameters are very effective in ranking the relevance of concepts in a single text, while they are less productive when KX is used for the purpose of terminology extraction from corpora.

Enrichment module based on external resources: After a domain corpus has been uploaded and all parameters have been set, the user will click the 'Extract relevant concepts' button to start the evaluation process. The key-concept list is disambiguated based on WordNet and is enriched by the synonyms extracted from the associated synsets. If available, some additional information of the matched synsets is also retrieved and displayed through the web interface (see first column of the table in Fig. 3), for example the WordNet definition, the hypernyms ("is a") and hyponyms, the SUMO Entry (Pease et al., 2002),[8] and the link to the corresponding Wikipedia page taken from the BABELNET resource (Navigli & Ponzetto, 2010). The current version of the tool uses only the synonym information for computing the evaluation metrics. However, we are investigating how to effectively exploit also the remaining additional information.

Evaluation module: The fourth module matches the uploaded ontology against the key-concept list enriched with the corresponding synonyms. The output of the module is a table where for each ranked key-concept, the following details are reported (see also the table shown in Fig. 3):

- its relevance with respect to the corpus;[9]
- any match between an ontology concept and the key-concept or one of the synonyms obtained from WordNet.

By clicking the 'Compute ontology metrics' button, the system computes the evaluation metrics described in Section 3. All metrics obtained are exported in a txt file, together with the matchings table obtained (stored in comma separated values form).

5. Typical application scenarios of a corpus-based ontology evaluation framework

The corpus-based ontology evaluation approach that we propose may support knowledge engineers in several ontology engineering related tasks: for instance, to rank available candidate ontologies according to their coverage of the domain of interest in order to evaluate the possibility of reusing some of them (*ontology ranking*), to understand whether an ontology reasonably covers the domain considered (*ontology corpus-adequacy evaluation*), or to understand which are domain-wise the most relevant concepts

[8] This information was extracted from http://sigmakee.cvs.sourceforge.net/sigmakee/KBs/WordNetMappings/.

[9] We recall that the relevance values are normalized, i.e. the highest relevance is set to one, and the others are computed proportionally accordingly.

Concepts extracted (Ordered by Relevance)	Relevance	100% matching	Synonym 100% matching
▶ activity	1.00000	X	
▶ attribute	0.88020		
sequence flow	0.71714	X	
▶ business process modeling notation	0.70216		
▼ task	0.49418	X	
▼ Wordnet			
▼ Synset_#00795720			
Wordnet Definition: any piece of work that is undertaken or attempted			
Is a: work			
Sumo Entry: IntentionalProcess			
▼ Synonyms			
undertaking			
project			
labor			
Hyponims: cinch, breeze, picnic, snap1, duck soup, child's play, pushover, walkover, piece of cake, adventure, escapade, risky venture, dangerous undertaking, assignment, baby, enterprise, endeavor, endeavour, labor of love, labour of love, marathon, endurance contest, no-brainer, proposition, tall order, large order, venture, Manhattan Project			
▶ Wikipedia Links			
▶ mapping	0.48253		
▶ flow	0.47920		
▶ message	0.43927	X	
▶ sub process	0.41265	X	
▶ gateway	0.39268	X	
▶ pool	0.30116	X	
message flow	0.27787	X	
▶ sequence	0.25790		
▶ expression	0.23461	X	
intermediate event	0.21963	X	
▶ token	0.21464		
end event	0.20799	X	
▶ gate	0.20799	X	
start event	0.20632	X	

Fig. 3. Evaluation output as displayed in the online interface. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AO-2012-0114.)

formalized in an ontology (*ranking of ontology concepts*). Below we describe them in more details, recalling to the reader that our approach is focused on terminological evaluation, and therefore it can be combined with complementary techniques for ontology evaluation focusing on more structural/formal aspects (see e.g. Gómez-Pérez, 2001; Brank et al., 2005), to provide a more complete framework for ontology evaluation.

5.1. Ontology ranking

State-of-the-art ontology modelling methodologies (see e.g. Gómez-Pérez et al., 2004) emphasize the importance of reusing publicly available ontologies, to reduce the modelling effort and speed up the ontology building process. However, in some situations, two or more potentially relevant ontologies may be available, and therefore it may be useful for the knowledge engineers to have a ranking of these ontologies according to their coverage of the domain of interest, in order to facilitate their comparison

and to reduce the effort of accurately inspecting all of them. The corpus-based ontology evaluation approach that we propose can efficiently and effectively support this task, as the modeller can compute the corpus-based ontology metrics proposed for each candidate ontologies, and sort them according to the metrics values returned.

5.2. Ontology corpus-adequacy evaluation

Understanding whether a given ontology adequately covers the domain of interest is a common and important issue when evaluating an ontology. For instance, a user (or a team of knowledge engineers and domain experts working collaboratively) formalizing a domain of interest may want to:

- periodically check whether the ontology under development is aligned with the standard terminology of the domain of interest (i.e. a sort of *ontology progress check*), or
- understand if a publicly available ontology is relevant and adequate for the domain to be modeled, in order to consider its possible adoption.

The corpus-based ontology evaluation approach that we propose supports this task by providing some objective numerical measures that can guide knowledge engineers in determining whether a given ontology reasonably covers the domain considered.

Actually, the output produced with our approach has also a constructive value: if many key-concepts returned from the corpus are not formalized in the ontology, the knowledge engineer can consider to enrich it with (a part of) these key-concepts.

5.3. Ranking of ontology concepts

Finally, another task supported by our approach is the evaluation of the relevance of the concepts in an ontology. Given that an ontology can include thousands of concepts, it is not always easy to understand if some of them are more relevant than others, thus representing a sort of "core" knowledge of the domain of interest (an information which usually is not explicitly encoded in ontologies). Indeed, users may find this information very important in order to:

- have a better understanding of the ontology, e.g. helping them in selecting the terms from which to start inspecting more in details the ontology, or
- plan, when building an ontology, which part of it should be first enriched with axioms and additional information.

The corpus-based ontology evaluation approach that we proposed can effectively support also this specific task: the user can take advantage of the relevance score associated with every key-concept extracted from the domain corpus. The rank of a key-concept that appears also in the ontology represents the importance of such ontology concept with respect to the domain.

6. Experimental evaluation

We performed some experiments to validate the corpus-based ontology evaluation approach and system that we propose. First of all, we introduce the corpora and ontologies considered throughout the experiments.[10]

[10] We remark that the system works with any OWL ontology and any corpus of digital text files.

6.1. Text Corpora

In the experiments performed, we considered the following two corpora.

BPMN Corpus: this corpus consists of the pages of the Business Process Modelling Notation (BPMN) Specification (OMG, 2008), i.e. the technical document describing one of the state-of-the-art (graphical) language for specifying business processes, BPMN. The BPMN Specification consists of 318 pages (approximately 93,000 words) presenting all the elements of BPMN, together with their attributes and properties;

Pizza Menus Corpus: this corpus consists of 50 pizza restaurant menus (approximately 22,000 words) collected over the Internet. These menus describe pizza types, ingredients, types of crust, details on sizes and prices, as well as information on additional products offered by the restaurants (e.g. beverages, sandwiches, and so on).

6.2. Ontologies

We considered four ontologies, covering different domains:

OntoBPMN (the BPMN Ontology): an OWL-DL formalization of BPMN (Ghidini et al., 2008). The ontology formally describes all BPMN elements, their attributes and the properties stating how the elements can be combined to form a structurally valid BPMN process representation. It was manually developed strictly following the BPMN Specification, i.e. the BPMN Corpus. The ontology comprises 116 concepts, and it is available for download at https://dkm.fbk.eu/index.php/BPMN_Ontology;

BMO (the Business Management Ontology): This ontology represents an integrated information model covering the business process design, project management, requirements management, and business performance management domains. In particular, the ontology allows to represent business processes, entities and objects. The ontology comprises 468 concepts, and it is available for download at http://www.bpiresearch.com/Resources/RE_OSSOnt/RE_BMO_DL/re_bmo_dl.htm;

Pizza Ontology: an ontology describing the domain of pizzas, including pizza types, pizza toppings and pizza bases. The ontology comprises 97 concepts, and it is available for download at http://www.co-ode.org/ontologies/pizza/2007/02/12/;

Food Ontology: an ontology describing the domain of food. The ontology comprises 136 concepts, and it is available for download at http://www.w3.org/TR/owl-guide/food.rdf.

Next, in Sections 6.3–6.5, we describe the experiments we performed and discuss the findings we concluded from them.

6.3. *Experiment* 1: *Investigating the ontology metrics for ontology ranking and corpus-adequacy evaluation*

The first experiment aims at investigating the behavior of the corpus-based ontology metrics proposed for supporting the tasks of ontology ranking and corpus-adequacy evaluation. The experiment comprises two parts, namely *Experiment* 1*a* and *Experiment* 1*b*: in each sub-experiment, we considered a domain corpus and three ontologies, and we computed the ontology metrics presented in Section 3 using our corpus-based ontology evaluation system.

6.3.1. *Experiment* 1*a*: BPMN Corpus *versus* OntoBPMN, BMO, *and* Pizza Ontology

We first performed a corpus-based evaluation of OntoBPMN, BMO, and Pizza Ontology against the BPMN Corpus. The choice of this combination of corpus/ontologies was not by chance. Indeed, we se-

Table 1
Results of Experiment 1a

	OntoBPMN	BMO	Pizza ontology
Concepts	116	468	97
Hits	58	16	0
Precision	0.50000	0.03419	0
Recall	0.11600	0.03200	0
$F1$	0.18831	0.03306	–
$Recall_w$	0.35375	0.13484	0
$F1_w$	0.41435	0.05455	–

lected one ontology which precisely formalizes the domain of interest (OntoBPMN, which was manually built starting from the BPMN Corpus), one ontology which describes a domain partially overlapping with the domain of interest considered (BMO), and one ontology formalizing a domain completely unrelated with the one covered by the BPMN Corpus (Pizza Ontology).

The extraction of key-concepts from the BPMN Corpus produced 500 entries[11] (222 of which enriched with domain or WordNet content). Then, we used MoKi to compute the evaluation metrics presented in Section 3. Table 1 reports for each ontology under evaluation the following information:

- *Concepts*: the number of concepts defined in the ontology;
- *Hits*: the number of concepts defined in the ontology matching the extracted key-concepts;
- *Precision*, *Recall*, $F1$, $Recall_w$ and $F1_w$.

6.3.2. Experiment 1b: Pizza Menus Corpus *versus* Pizza Ontology, Food Ontology *and* OntoBPMN

In the second part of Experiment 1, we performed a corpus-based evaluation of Pizza Ontology, Food Ontology, and OntoBPMN against the Pizza Menus Corpus. Again, the choice of this combination of corpus/ontologies was made on purpose: we selected one ontology supposed to cover particularly well the domain considered (Pizza Ontology), one ontology which describes a domain more general than the domain of interest considered (Food Ontology), and one ontology formalizing a domain completely unrelated to the one covered by the Pizza Menus Corpus (OntoBPMN).

The extraction of key-concepts from the BPMN Corpus produced 202 entries (150 of which enriched with domain specific or WordNet content). Similarly to Experiment 1a, we computed the ontology metrics presented in Section 3. Table 2 reports the results obtained for each ontology under evaluation.

6.3.3. Findings of Experiment 1

The purpose of Experiment 1 was to evaluate the capability of our approach in (i) supporting the ranking of ontologies according to a reference corpus, and (ii) determining whether an ontology adequately covers a given domain from a terminological point of view. Concerning the ontology ranking task, both $F1$ and $F1_w$ return the same ontology ranking, on both corpora:

– BPMN Corpus
  * $F1$: 1. OntoBPMN (0.18831), 2. BMO (0.03306) and 3. Pizza Ontology (Undefined)
  * $F1_w$: 1. OntoBPMN (0.41435), 2. BMO (0.05455) and 3. Pizza Ontology (Undefined)

[11] We recall that the system discards all key-concepts having a normalized relevance lower than 0.01.

Table 2
Results of Experiment 1b

	Pizza ontology	Food ontology	OntoBPMN
Concepts	97	136	116
Hits	26	2	0
Precision	0.26804	0.01471	0
Recall	0.12871	0.00990	0
$F1$	0.17391	0.01183	–
$Recall_w$	0.23781	0.00855	0
$F1_w$	0.25202	0.01082	–

– Pizza Menus Corpus
  * $F1$: 1. Pizza Ontology (0.17391), 2. Food Ontology (0.01183) and 3. OntoBPMN (Undefined)
  * $F1_w$: 1. Pizza Ontology (0.25202), 2. Food Ontology (0.01082) and 3. OntoBPMN (Undefined)

The rankings outputted by the system are quite significant, as the difference between a value and the following one in the ranking is extremely sharp. For instance, $F1_w$ of OntoBPMN on the BPMN Corpus is approximately 8 times bigger than the one for BMO, while $F1_w$ of Pizza Ontology on the Pizza Menus Corpus is approximately 23 times bigger than the one for Food Ontology. The rankings confirm exactly the criteria we applied for selecting the ontologies to be compared, thus suggesting that the use of $F1$ and $F1_w$ provides effective support in ranking ontologies according to a given corpus.

Concerning the evaluation of ontology corpus-adequacy, we performed Experiment 1 choosing, for each corpus considered, one "optimal" ontology for the domain covered by the corpus, i.e. OntoBPMN for the BPMN Corpus, and Pizza Ontology for the Pizza Menus Corpus. Indeed, OntoBPMN was manually built by formalizing the BPMN Corpus, while the Pizza Ontology was built independently of the Pizza Menus Corpus. In view of these facts, the results obtained allow us to propose some guidelines for determining whether an ontology adequately represents a given reference corpus. Given an ontology and a domain corpus, we can conclude that the ontology provides an adequate terminological coverage of the domain if:

– $F1$ is greater or equal than 0.15, or
– $F1_w$ is greater or equal than 0.25.

6.4. *Experiment* 2: *Showing the effect of key-concept relevance on the ontology metrics proposed*

The goal of Experiment 2 is to show the impact of considering a list of key-concepts ranked by corpus relevance in the evaluation.

We perform Experiment 2 using the BPMN Corpus and OntoBPMN. Many concepts in OntoBPMN match with top-ranked key-concepts obtained from the corpus (37 hits among the top 100 key-concepts), while there are only 21 hits among the key-concepts ranked from position 101–500. This means that OntoBPMN captures many concepts that are highly relevant in the domain, in addition to other concepts that have lower relevance. $Recall_w$ and $F1_w$ have been introduced to represent exactly this aspect: in contrast to standard *Recall* and $F1$, they return higher values if the hits in the evaluated ontology are highly relevant concepts in the domain.

In order to show this, we evaluated the ontology OntoBPMN against (a) the ranked list of key-concepts returned by the system as is (*Standard Relevance Values*) and (b) the reversed list, created by assigning

Table 3
Comparison of evaluation metrics considering the ranked list in standard and reversed order

	Standard relevance values	Reversed relevance values
Recall	0.11600	0.11600
$F1$	0.18831	0.18831
$Recall_w$	0.35375	0.06924
$F1_w$	0.41435	0.12163

to the ith key-concept of the list the relevance value of the $(500 - i)$th key-concept of the list (*Reversed Relevance Values*).

The rationale behind this "artificial" configuration is to simulate the situation of having two ontologies of similar size, both matching the same number of key-concepts from the list, but one ontology matching many key-concepts having relatively high relevance value, i.e. the most relevant key-concepts in the domain corpus according to the system, while the other one matching many key-concepts having relatively low relevance value. Ideally, in such situation, the former ontology would be preferable to the latter, as it covers a more relevant part of the domain.

6.4.1. Findings of Experiment 2

With this specific configuration, the standard $F1$ does not reveal to be particularly useful, as its value is exactly the same in both cases, as reported in Table 3. On the contrary, $F1_w$ discriminates well between the two situations, returning a higher value in the case of the standard ordering (as expected).

The experiment shows the capability of $Recall_w$ and $F1_w$ to provide a more accurate ontology evaluation than standard *Recall* and $F1$ by exploiting the relevance value associated to the key-concepts extracted from the corpus. This suggests that these metrics may be applied to rank ontologies also in situations where $F1$ does not provide a clear discrimination between the candidates. Furthermore, the comparison between $F1$ and $F1_w$ obtained by matching an ontology against a reference corpus provides us with further useful information: if $F1_w$ is greater than $F1$, then the ontology covers more concepts with relatively high relevance than those with lower relevance. If not, the contrary holds.

6.5. Experiment 3: *Ranking the* OntoBPMN *concepts*

The third analysis that we report aims at showing the capability of our corpus-based ontology evaluation approach to rank the concepts in a given ontology according to their relevance with respect to the reference corpus. The experiment involved the BPMN Corpus and OntoBPMN ontology. We considered the 15 top-ranked key-concepts of the BPMN Corpus that match concepts in OntoBPMN. The list of such concepts (together with their normalized relevance) is shown in Table 4. Note that these concepts are among the 21 top-ranked key-concepts of the whole BPMN Corpus.

6.5.1. Findings of Experiment 3

The analysis of this ordered list shows that 12 out of 15 of the elements are the basic (graphical) elements of most BPMN business processes (see e.g. Muehlen & Recker, 2008; OMG, 2008):

- *business process diagram* identifies the whole process;
- *flow objects* represent the main graphical elements to define the behaviour of a process; according to the BPMN Specification, flow objects are organized in 3 sub-categories: events (representing something that happens), *activities* (work to be performed), or *gateways* (control flow elements);
- an *activity* is further specialized in *sub-process* (complex activity) and *task* (atomic activity);

Table 4
Automatic corpus-based ranking of OntoBPMN concepts

Ontology concept	Relevance
Sequence flow	0.77416
Activity	0.35625
Message flow	0.35448
Attribute	0.31357
Sub-process	0.29401
Task	0.17605
Intermediate event	0.15649
End event	0.14878
Start event	0.14641
Gateway	0.13989
Business process diagram	0.10966
Pool	0.10729
Web service	0.08477
Expression	0.08358
Flow object	0.08121

- *start event*, *end event*, and *intermediate event* are the main types of event;
- *sequence flow* and *message flow* are the main types of connecting objects, i.e. the graphical elements used to connect the different flow objects in a process.

As for the three remaining concepts (*attribute*, *web service*, *expression*), they do not refer to basic graphical elements, although they are needed to represent more complex and detailed business processes. Indeed, Muehlen and Recker (2008) presented an analysis of the most common subsets of BPMN elements used in actual business process diagrams. The analysis revealed that the subsets of most used elements in BPMN process diagrams are:

- *task* and *sequence flow* in 97% of cases;
- *task*, *sequence flow*, *start event*, and *end event* in 54% of cases;
- *task*, *sequence flow*, *start event*, *end event*, and *gateway* in 22% of cases;
- *task*, *sequence flow*, *start event*, *end event*, *gateway*, and *pool* in 10% of cases.

The elements of these subsets all appear among the first 6, 9, 10, 12 (respectively) key-concepts presented in Table 4. Therefore, under the underlying assumption that the most relevant elements of BPMN are also the most used ones in practice, we can fairly conclude that our approach for corpus-based ontology evaluation contributes to the identification of the most relevant key-concepts represented in the ontology.

7. Limitations of the evaluation approach

Despite the encouraging results shown by our experimental evaluation, we are aware that the work described in this paper presents some limitations that we discuss in detail in this section.

As already remarked, the framework currently deals with the *terminological* evaluation of ontologies, comparing the terms used as concepts in an ontology with the terms characterizing a text corpus. The work could be extended by considering additional ontological entities that can be identified in the corpus, e.g. properties and individuals. We are extending the proposed framework in this direction, by enhancing

the capabilities of the key-concept extraction phase. Consequently, we will adapt the evaluation metrics to handle the additional entities considered.

The current approach does not provide any feedback about the correctness and completeness of the axiomatization of a domain encoded in the ontology (what we refer to as *semantic level* in Section 1). In other words it does not provide a measure of the completeness and correctness of the OWL axioms contained in an ontology. In standard ontology engineering methodologies, this aspect is evaluated by comparing what can be inferred from the ontology from a set of *competency questions* (Grüninger & Fox, 1995), i.e. a list of questions that a knowledge base based on the ontology should be able to answer. The extension of our approach to cover also this aspect would imply the capability of automatically extracting competency questions from the corpus, which is analogous to the operation of automatic ontology construction. However, the automatic extraction of axioms from a corpus of document is still a challenging research direction, offering not very consolidated results. In most of the cases, the axioms automatically extracted need to be carefully checked and revised by experts, and therefore they cannot be used as a gold standard for the evaluation of an ontology. Furthermore, it can be argued that if we were able to automatically generate *gold-standard ontologies* from corpora, then the need to evaluate ontologies would be put back into perspective.

Our approach includes a disambiguation step, in which each key-concept is unambiguously linked to a single WordNet synset. This feature could be further exploited to integrate in our evaluation workflow the additional semantic information encoded in WordNet, such as WordNet hierarchical information, the nouns in the glosses, and suitable logical translations of the glosses (Harabagiu et al., 1999). This improvement will be investigated in our future research activities. We will also devote some effort to the analysis and evaluation of other disambiguation strategies.

The evaluation framework relies on a key-concept extraction phase which is performed in a fully automatic and unsupervised manner. Although the system exploits state of the art key-concept extraction techniques, we note that the list of the terms automatically obtained from the corpus may not coincide with the set of concepts characterizing the domain of interest. This because the extraction phase, together with key-concepts which characterize the domain of interest, may return key-concepts that are not (or marginally) in the domain, or terms which are not proper key-concepts (due to errors). We are aware that this fact may have an impact on the results obtained computing the metrics that we propose. For example, an ontology with a broad scope may show better evaluation values than a smaller and more focused ontology, even if the latter represents better than the former the terminology used in the domain described by the corpus. This should be considered also when interpreting the results obtained computing the metrics that we propose. In fact, although optimal values according to the metrics are those close to 1.0, they are practically beyond one's reach. As an example, we recall that $F1_w$ obtained for the OntoBPMN against the BPMN Corpus is 0.46363, although the ontology was manually built exactly from that corpus. Still, the obtained values can be used as useful indices (if not full measures) of the terminological coverage of an ontology, and allow for meaningful comparisons between different ontologies.

8. Conclusions and future work

In this work, we presented a novel framework for the terminological evaluation of ontologies against a domain of interest represented through a text corpus. The framework is based on the automatic extraction of domain-specific key-concepts. Each key-concept extracted has an associated relevance value, which represents how relevant the key-concept is in the domain of interest.

Our corpus-based ontology evaluation framework matches the concepts in the ontology with the domain key-concepts extracted from the corpus, exploiting additional linguistic resources like WordNet synonyms to obtain a more accurate matching. Based on this matching, several metrics can be computed to obtain some objective measures of whether a given ontology adequately covers the domain of interest. Beside standard metrics borrowed from the Information Retrieval field (*Precision*, *Recall*, and $F1$), we introduced two new weighted metrics ($Recall_w$ and $F1_w$) that take into account not only the matches between the domain key-concepts and the ontology, but also the relevance of the matched key-concepts.

We performed a thorough experimental analysis of our approach, showing that these metrics capture well the adequacy of an ontology with respect to a domain, and allow users to *i)* effectively and efficiently rank candidate ontologies according to how they terminologically cover a given domain, and *ii)* understand which are domain-wise the most relevant concepts formalized in an ontology.

Furthermore, all metrics and experiments presented in this paper can be reproduced because we have made available a system for ontology building, extension and evaluation, which allows different users to work collaboratively at the selection and evaluation of ontologies. The system can be used online through a user-friendly interface, where evaluations and analyses similar to the ones presented in the previous sections can be easily performed. To the best of our knowledge, this is the first (collaborative) system of this kind made publicly available.

In future, we plan to further improve different aspects of our approach. First, we will consider WordNet hypernyms during the matching process between ontology and key-concepts, in order to evaluate coverage in a more flexible way. Also, we will take into account sub/super-string matches. Besides, we will further exploit WordNet information by matching WordNet relations to the ontology structure, so that also the structural aspects of the ontology can be included in the evaluation. With this respect, we can build on some recent proposals advanced in the literature (e.g. Bolotnikova et al., 2011).

Acknowledgements

The work described in this paper has been partially funded by the European Commission under the contract number FP7-248594, PESCaDO project.[12]

References

Blei, D.M., Ng, A.Y. & Jordan, M.I. (2002). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4,5), 993–1022.

Bolotnikova, E.S., Gavrilova, T.A. & Gorovoy, V.A. (2011). To a method of evaluating ontologies. *Journal of Computer and Systems Sciences International*, *50*(3), 448–461.

Brank, J., Grobelnik, M. & Mladenic, D. (2005). A survey of ontology evaluation techniques. In *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005)*, Ljubljana, Slovenia (pp. 166–170). Citeseer.

Brewster, C., Alani, H., Dasmahapatra, A. & Wilks, Y. (2004). Data driven ontology evaluation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal.

Buitelaar, P., Cimiano, P. & Magnini, B. (Eds.) (2005). *Ontology Learning from Text: Methods, Evaluation and Applications* (Vol. 123). Amsterdam: IOS Press.

Cui, H. (2010). Competency evaluation of plant character ontologies against domain literature. *Journal of the American Society for Information Science and Technology*, *61*(6), 1144–1165.

Dellschaft, K. & Staab, S. (2006). How to perform a gold standard based evaluation of ontology learning. In *Proceedings of ISWC-2006 International Semantic Web Conference*, Athens GA, USA.

[12]Personalized Environmental Service Configuration and Delivery Orchestration.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: MIT Press.
Frantzi, K., Ananiadou, S. & Mima, H. (2000). Automatic recognition of multi-word terms: The C-value/NC-value. *Journal of Digital Libraries*, *3*(2), 115–130.
Ghidini, C., Rospocher, M. & Serafini, L. (2008). A formalisation of BPMN in description logics. Technical Report TR 2008-06-004, FBK-irst. Available at: https://dkm.fbk.eu/index.php/BPMN_Related_Resources.
Ghidini, C., Rospocher, M. & Serafini, L. (2010). MoKi: A wiki-based conceptual modeling tool. In *CEUR Workshop Proceedings*, ISWC 2010 Posters & Demonstrations Track: Collected Abstracts (Vol. 658, pp. 77–80), Shanghai, China.
Gómez-Pérez, A. (2001). Evaluation of ontologies. *International Journal of Intelligent Systems*, *16*(3), 391–409.
Gómez-Pérez, A., Fernández-López, M. & Corcho, O. (2004). *Ontological Engineering*. Berlin: Springer.
Grüninger, M. & Fox, M.S. (1995). *Methodology for the Design and Evaluation of Ontologies* (Vol. 95, pp. 6.1–6.10). Menlo Park, CA, USA: AAAI Press.
Harabagiu, S.M., Miller, G.A. & Moldovan, D.I. (1999). WordNet 2 – A morphologically and semantically enhanced resource. In *Proceedings SIGLEX 1999* (pp. 1–8). Maryland, MD, USA: University of Maryland.
Hulth, A. (2004). Combining machine learning and NLP for automatic keyword extraction. PhD thesis, Stockholm University.
Jones, M. & Alani, H. (2006). Content-based ontology ranking. In *Proceedings of the 9th International Protege Conference*, Stanford, CA, USA (pp. 23–26).
Lee, G., Mariam, T. & Ahmad, K. (2005). Terminology and the construction of ontology. *Terminology*, *11*(1), 55–81.
Liddle, S.W., Hewett, K.A. & Embley, D.W. (2003). An integrated ontology development environment for data extraction. In *Proceedings of Information Systems Technology and Its Applications, International Conference (ISTA)*, Kharkiv, Ukraine (pp. 21–33).
Lozano-Tello, A. & Gómez-Pérez, A. (2004). ONTOMETRIC: A method to choose the appropriate ontology. *Journal of Database Management*, *15*(2), 1–18.
Maedche, A. (2002). *Ontology Learning from the Semantic Web*. Kluwer International Series in Engineering and Computer Science. Norwell, MA, USA: Kluwer Academic Publishers.
Maedche, A. & Staab, S. (2002). Measuring similarity between ontologies. In *Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW 2002)*. Lecture Notes Computer Science (Vol. 5074, pp. 251–263). Berlin, Germany: Springer.
Muehlen, M. & Recker, J. (2008). How much language is enough? Theoretical and practical use of the business process modeling notation. In *Proceeding CAiSE '08 Proceedings of the 20th international conference on Advanced Information Systems Engineering* (pp. 465–479). Berlin, Germany: Springer.
Navigli, R. & Ponzetto, S.P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden.
Navigli, R. & Velardi, P. (2004). Learning domain ontologies from document warehouses and dedicated websites. *Computational Linguistics*, *30*, 151–179.
OMG (2008). Business process modeling notation, v1.1. Available at: www.omg.org/spec/BPMN/1.1/PDF.
Pease, A., Niles, I. & Li, J. (2002). The suggested upper merged ontology: A large ontology for the semantic web and its applications. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, Edmonton, Canada. Menlo Park, CA, USA: AAAI.
Pedersen, T., Banerjee, S. & Patwardhan, S. (2005). Maximizing semantic relatedness to perform word sense disambiguation. Technical Report UMSI 2005/25, University of Minnesota Supercomputing Institute.
Pianta, E. & Tonelli, S. (2010). KX: A flexible system for Keyphrase eXtraction. In *Proceedings of SemEval 2010, Task 5: Keyword Extraction from Scientific Articles*, Uppsala, Sweden.
Porzel, R. & Malaka, R. (2004). A Task-Based Approach for Ontology Evaluation. In *Proceedings of the ECAI Workshop on Ontology Learning and Population*, Valencia, Spain.
Tonelli, S., Rospocher, M., Pianta, E. & Serafini, L. (2011). Boosting collaborative ontology building with key-concept extraction. In *Proceedings of the Fifth IEEE International Conference on Semantic Computing*, Stanford, CA, USA (pp. 316–319).
Witten, I., Paynter, G., Frank, E., Gutwin, C. & Nevill-Manning, C. (1999). KEA: Practical Automatic Keyphrase Extraction. In *Proceedings of the Fourth ACM Conference on Digital Libraries*, Berkeley, CA, USA (pp. 254–256).
Wong, W., Liu, W. & Bennamoun, M. (2007). Determining termhood for learning domain ontologies using domain prevalence and tendency. In *AusDM'07, Proceedings of the Sixth Australasian Conference on Data Mining and Analytics* (Vol. 70, pp. 47–54). Australia: Australian Computer Society, Inc.
Yao, L., Divoli, A., Mayzus, I., Evans, J.A. & Rzhetsky, A. (2011). Benchmarking ontologies: Bigger or better? *PLoS Computational Biology*, *7*, e1001055.