# Language and domain aware lightweight ontology matching

Web Semantics: Science, Services and Agents on the World Wide Web ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

# Language and domain aware lightweight ontology matching

Gábor Bella [a,*], Fausto Giunchiglia [a], Fiona McNeill [b]

[a] University of Trento, via Sommarive 5, 38123 Trento, Italy
[b] Heriot-Watt University, Edinburgh EH14 4AS, Scotland, United Kingdom

ARTICLE INFO

*Article history:*
Received 24 August 2016
Received in revised form 15 March 2017
Accepted 26 March 2017
Available online xxxx

*Keywords:*
Cross-lingual matching
Multilingual matching
Domains
Ontology matching
Semantic matching
Machine translation

ABSTRACT

Concepts and relations in ontologies and in other knowledge organisation systems are usually annotated with natural language labels. Most ontology matchers rely on such labels in element-level matching techniques. State-of-the-art approaches, however, tend to make implicit assumptions about the language used in labels (usually English) and are either domain-agnostic or are built for a specific domain. When faced with labels in different languages, most approaches resort to general-purpose machine translation services to reduce the problem to monolingual English-only matching. We investigate a thoroughly different and highly extensible solution based on *semantic matching* where labels are parsed by multilingual natural language processing and then matched using language-independent and domain aware background knowledge acting as an interlingua. The method is implemented in NuSM, the language and domain aware evolution of the SMATCH semantic matcher, and is evaluated against a translation-based approach. We also design and evaluate a fusion matcher that combines the outputs of the two techniques in order to boost precision or recall beyond the results produced by either technique alone.


## 1. Introduction

Ontologies and other knowledge organisation systems, while usually serving a purpose of standardisation or generalisation, stem from local needs and practices. By *local* we understand *within an administrative unit* such as a country or a region as well as *within an application domain* such as medicine or transport. Accordingly, ontologies tend to target specific domains and the labels annotating their elements – concepts, relations, metadata – tend to be expressed in the local language. This is especially true for *lightweight ontologies* [1]: classification hierarchies, taxonomies, and other tree-structured data schemas widely used around the world as simple, well-understood, semi-formal resources for knowledge organisation. Such resources often play normative roles on the national level in public services, industry, or commerce, as a means for classification (of documents, books, open data, commercial products, web pages, etc.) as well as being sources of shared vocabularies for actors cooperating in a given domain.

*Ontology matching* [2] is a process that creates and maintains alignments between elements of two ontologies covering overlapping areas of knowledge. We define *language aware* or *multilingual* matching as a type of ontology matching where a multilingual setting is explicitly assumed, i.e., the matcher is capable of dealing with ontologies expressed in multiple languages. Likewise, we define *domain aware* matching as capable of dealing with domain-specific knowledge and domain terms with specialised meanings.

Activities on supra-national levels such as international trade and mobility need to rely on the interoperability and integration of knowledge organisation resources across countries, languages, and sometimes across domains. *Cross-lingual matching* is a specific case of language aware matching when ontologies in different languages need to be aligned. Likewise, *cross-domain matching* is used to match ontologies pertaining to different domains of knowledge. An example of a simultaneously cross-lingual and cross-domain matching problem is the case of *cross-border emergency response* where responders from different countries and from different domains (geography, geology, medicine, police, military, transportation, etc.) need to share data. In [3] we apply the domain aware matching approach presented in this paper to this particular use case.

State-of-the-art cross-lingual matchers invariably use translation-based techniques – most often online machine translation services from Microsoft or Google – in order to reduce the problem of multilingualism to the well-researched problem of monolingual English-to-English matching [2,4–8, p. 105]. With the constant improvement of such services, translation-based matchers are able to provide usable results and are able to deal with a wide range of languages. State-of-the-art machine translators today mainly use statistical methods and are trained on large amounts of *bilingual parallel or comparable corpora* for each language pair they support.

* Corresponding author.
*E-mail address:* gabor.bella@unitn.it (G. Bella).

http://dx.doi.org/10.1016/j.websem.2017.03.003
1570-8268/

A known problem of statistical machine translation, however, is the decrease of translation accuracy on corpora significantly different from those on which the system was trained. This typically happens on domain classifications and ontologies that contain specialised terminology. The adaptation of a statistical system to a new domain requires re-training on corpora extended with a significant amount of domain-specific text (ideally bilingual parallel corpora that are hard to find). At the same time, the systems typically used by ontology matchers are on-line commercial services (such as Bing and Google Translate) that, while offering the best available translation quality, are not adaptable or customisable by the user.

The shortness of labels typically found in ontologies is another difficulty that state-of-the-art approaches face, as the sparseness of textual context within labels makes the translation task more error-prone. Furthermore, the often non-standard orthography and syntax of ontology labels – that we described in [9] as a form of specialised *block language*– makes label parsing even harder.

In this paper we introduce a different approach to language and domain aware matching, so far hardly investigated, that does not rely on external translation tools. The method is based on combining two types of resources: on the one hand, multilingual natural language processing tools that are adapted to the block language of structured data and, on the other hand, off-line *multilingual lexical databases* connecting words and expressions of natural language to language-independent but domain-aware meanings.

The work described in this paper is motivated by the following considerations. Firstly, while both approaches evoked above are resource-intensive, the types of resources they feed on are markedly different: on the one hand, machine translation requires large amounts of bilingual parallel or comparable corpora relevant to the target domain, on the other hand, our approach uses lexical, terminological, and NLP resources for each supported language. In both cases, a wide range of resources are already available on the web. Based on their availability and conditions of use, for specific use cases one approach or the other may be more cost-effective or faster to implement. Our knowledge-based label matching approach can thus be seen as an alternative when no good-quality language or domain-specific machine translator is available. Secondly, we are interested in comparing the strengths and weaknesses of the two approaches, which turn out to be rather complementary. Our evaluations use two machine translation systems: Google Translate, currently the best available online translator, and Apertium, which is free and can also be used off-line. We conduct evaluations on three language pairs: English–Spanish, English–Italian, and Spanish–Italian. Finally, based on the complementarity of the two approaches we investigate the idea of combining them – using multilingual lexical resources on the one hand and machine translation on the other hand – into a single matcher. The resulting system, as demonstrated by our evaluation results, clearly outperforms either method alone.

The result of our work is implemented in *NuSMATCH* (NuSM for short), an upcoming release of the open-source SMATCH system [10] with built-in capabilities for language and domain aware matching.

The rest of the paper is organised as follows. After a brief reminder of semantic matching in Section 2, Section 3 introduces language and domain aware matching. Section 4 presents the core multilingual and multidomain resource serving as background knowledge for our matcher. Sections 5 and 6 deal with interpreting labels, explaining how NuSM performs language aware parsing and domain aware sense disambiguation, respectively. Section 7 describes our cross-lingual label matching techniques. Section 8 provides mechanisms for extending NuSM by languages and domain terms. Section 9 proposes a fusion matcher that combines NuSM with translation-based matching. Section 10 provides evaluations for NuSM both in comparison to and in combination with translation-based techniques. Section 11 presents related work and, finally, Section 12 draws the overall conclusions.

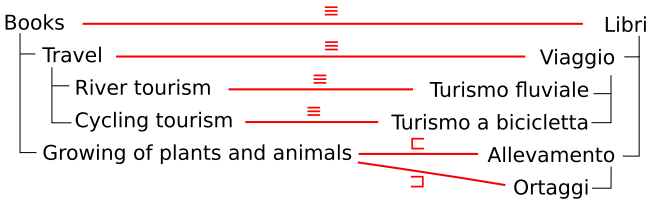



**Fig. 1.** Example English and Italian classifications of books, with some example mapping relations.

## 2. Semantic matching

In this section we provide an overview of the notions of *lightweight ontology*, *semantic matching*, and of the principal aspects of language and domain awareness in NuSM.

### 2.1. Principles of semantic matching

NuSM is designed as a multilingual and domain aware extension of the SMATCH (English-only) semantic matcher [10,11]. SMATCH was specifically designed for matching semi-formal knowledge organisation schemes such as classification hierarchies (as opposed to formal ontologies expressed, e.g., in OWL) that we believe are the main subject of most real-world multilingual matching applications. As shown in [1], such classifications typically have the following properties:

- a tree structure;
- nodes are expressed as short and well-formed natural language labels;
- classification semantics are implied for nodes and edges.

As an example of classification semantics, in a classification of newspaper articles the extension of a node *Literature* are articles about literature, while a node *Italy* under *Literature* classifies articles on Italian literature.

Matching is qualified as *semantic* for three reasons:

- it is performed using logical inference on propositional description logic formulas that capture the meanings of natural-language classification labels in a formal way;
- atoms of the formulas are concepts and are not based on the surface forms of words;
- the ontology node mappings output by the matcher are 'meaningful' description logic relations of equivalence, subsumption, and disjointness as opposed to similarity scores.

These operators are important in several real-world matching tasks such as data integration or query answering. For example, in a schema matching operation between query and database attributes, in response to a query on a *'phone_number'* a database may return a *'mobile_phone_number'* (which is subsumed by the meaning of the query) but the contrary may not be correct. See Fig. 1 for an example of multilingual trees as input and of mappings as output of NuSM.

### 2.2. The semantic matching process

SMATCH matches its two input trees in a four-step process where the two first steps consist of formalising each input tree into a so-called *lightweight ontology* [1] while the two last steps perform the actual matching using background knowledge:

Please cite this article in press as: G. Bella, et al., Language and domain aware lightweight ontology matching, Web Semantics: Science, Services and Agents on the World Wide Web (2017), http://dx.doi.org/10.1016/j.websem.2017.03.003.

1. computation of label formulas;
2. computation of node formulas;
3. label matching using background knowledge axioms;
4. matching of the two trees.

Below we provide a brief overview of what is common between SMATCH and NuSM, while Section 3 and the rest of the paper present the new aspects of NuSM. For a more detailed presentation of semantic matching and the original SMATCH tool, we refer the reader to [11] and [10].

*Step 1* processes natural language labels in the two input trees and generates their formal semantic representations in the form of *propositional description logic formulas*. Constants in label formulas are atomic concepts from a background lexical database (Princeton WordNet in the case of SMATCH) while operators are conjunction, subjunction, and negation.

Formula computation is essentially a natural language processing task where concepts corresponding to lemmas (i.e., dictionary forms) of words are retrieved from the lexical database while operators and bracketing are computed according to the syntactic structure of the label. For example, in Fig. 1, for the English label *'Growing of plants and animals'* the formula $growing \sqcap (plant \sqcup animal)$ is computed where *'growing'*, *'plant'*, and *'animal'* are concepts, the word *'of'* is interpreted as a conjunction and the word *'and'* as a disjunction (since the meaning of the coordinating conjunction 'and' is typically the union and not the intersection of its arguments: in reality, the node classifies books either about plants *or* about animals).

*Step 2* incorporates further contextual information into the logical formulas computed from the tree structure. The operation consists of extending label formulas into *node formulas* by adding contextual information from other nodes in the tree. According to classification semantics, a node formula is computed by taking the conjunction of its label formula with the label formulas of all of its ancestors. For *'Growing of plants and animals'* in Fig. 1, the node formula $book \sqcap growing \sqcap (plant \sqcup animal)$ is computed, formalising the fact that this node classifies objects that are books and are about growing either plants or animals.

*Step 3* collects axioms relevant to the ontologies being matched from outside knowledge resources. For each meaning of each word in each label of the source tree, this step retrieves all ontological relations that hold between it and each meaning of each word in each label of the target tree. In SMATCH, WordNet is used as the principal knowledge base with meanings represented by synsets. In the absence of WordNet relations for a pair of synsets and also for words entirely missing from WordNet (out-of-vocabulary words), SMATCH uses string similarity and gloss matching techniques.

*Step 4* performs the matching task by running a SAT solver on pairs of source–target node formulas $(f_S, f_T)$, computed in step 2 and complemented by corresponding axioms retrieved in step 3. If a pair turns out to be related by one of three relations: *equivalence* $f_S \leftrightarrow f_T$, *implication* $f_S \leftarrow f_T$ or $f_S \rightarrow f_T$, or *negated conjunction* $\neg(f_S \wedge f_T)$ then the mapping relation equivalence, subsumption, or disjointness is returned as a result, respectively. If none of the above holds, a no-match relation is returned.

## 3. Language and domain aware semantic matching

By *language and domain aware matching* we understand a process where the language and the domain(s) of the input are *explicit parameters* instead of being implicitly assumed and, furthermore, where the matcher is *extensible* by resources and tools specific to the languages and domains to be supported. We achieve extensibility by means of the following two components, as depicted in Fig. 2:

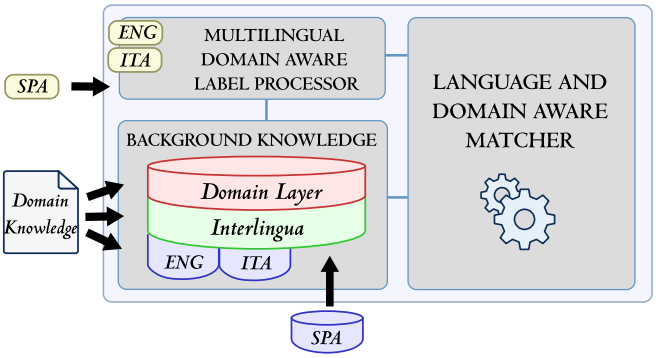



**Fig. 2.** Main components and extensibility features of the matcher: extension by a new language (here: Spanish) or by domain knowledge.

- an extensible *multilingual and domain aware knowledge base* as principal source of background knowledge;
- a *multilingual and domain aware label processor*, extensible to new languages in a straightforward manner.

The *background knowledge*, described in detail in Section 4, is a knowledge base containing *lexical databases* (i.e., wordnet) for each language supported, a language-independent ontology of concepts serving as an *interlingua*, and a set of *domains* mapped to each of those concepts providing a domain-based categorisation of lexical meaning. Extensibility by a new language is achieved by plugging in a corresponding lexical database while the rest of the knowledge base is only minimally changed if at all. Extensibility by domain-specific terminology is achieved by adding new terms to an existing lexical database and to the corresponding concepts to the interlingua, mapped to the appropriate domain.

*Label processing* consists of a language aware *label parsing* step (cf. Section 5) followed by a language-independent and domain aware *label disambiguation* step (Section 6).

*Label parsing* is a multilingual natural language processing task optimised to the language of lightweight ontology labels and is extensible by language-specific NLP components. Itself consists of the following substeps:

1. *language detection* that makes the language of each input tree explicit,
2. *computation of formula structure* that parses the label using syntactic NLP techniques partly generalised and partly adapted to each language supported,
3. *computation of atomic concepts* that formalises meaningful words in the label as *language-independent concepts*.

*Domain aware label disambiguation* serves the purpose of reducing the number of possible meanings of ambiguous words in labels using a domain-driven word sense disambiguation technique. We thus assume that the domain specificity of labels is essentially a lexical phenomenon. Disambiguation is performed in two, entirely language-independent substeps:

1. *domain detection* makes the domain(s) of the input ontologies explicit,
2. *domain-based sense disambiguation* is performed as a downstream NLP operation.

We describe these components in detail in the following sections.

## 4. Language and domain aware background knowledge

Core to our approach is a *multilingual and multidomain knowledge resource* that we use as background knowledge for matching. It consists of three layers (Fig. 3):

Please cite this article in press as: G. Bella, et al., Language and domain aware lightweight ontology matching, Web Semantics: Science, Services and Agents on the World Wide Web (2017), http://dx.doi.org/10.1016/j.websem.2017.03.003.

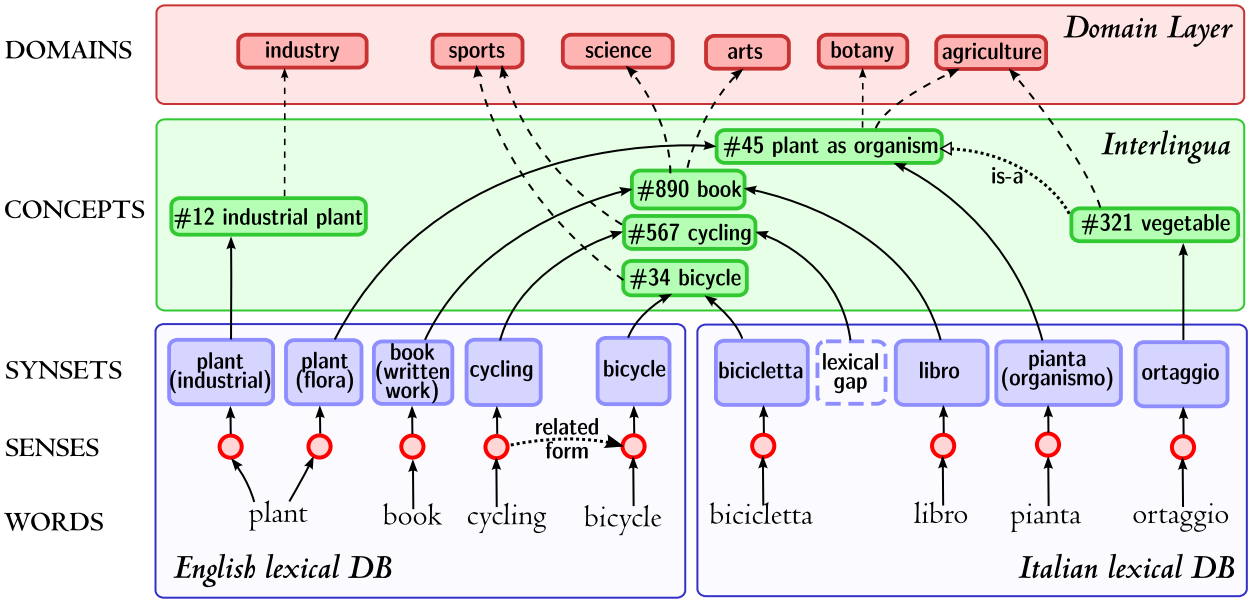



**Fig. 3.** Example English and Italian lexical databases with the interlingua layer providing language interoperability and the domain layer providing domain categories for concepts.

1. a lower layer of pluggable language-specific *lexical databases* that are WordNet-like lexical-semantic resources;
2. the *interlingua*: a language-independent but domain aware ontology of concepts where each concept is linked to its corresponding lexical entries in each language;
3. the *domain layer* that provides explicit domain information by associating interlingua concepts to domain categories.

The architecture of *lexical databases* is similar to that of Princeton WordNet [12], consisting of *lemmas* (i.e., dictionary forms of words of a language) associated to formally defined *word senses*. Synonymous senses are grouped together in synonym sets or *synsets*. Both senses and synsets are interconnected by lexical-semantic relations. Synsets represent an abstraction from the language-specific lexicon towards units of lexical meaning. Each lexical database is extensible by language-specific domain terms as described in Section 8.

The principal role of the language-independent *interlingua* layer is to serve as a bridge between language-specific lexical databases. Each synset in a lexical database is mapped to precisely one concept representing the corresponding language-independent unit of meaning. The opposite is not necessarily true. If a lexical database is incomplete or if a concept does not have a lexicalisation in a language, it will not be mapped to a synset in that language. This phenomenon is known as *lexical gap*. For example, there is no Italian word for *cycling*, which is paraphrased as 'andare a bicicletta', 'to go on a bicycle'. In this case, the concept of *cycling* is connected to an English synset but not to any Italian synset as it has no lexicalisation in that language.

The interlingua acts as an interoperability layer across language-specific lexical entries, a feature that we use for cross-lingual matching. We consider two lexemes (words or expressions) in two different languages to be equivalent (to 'mean the same') if their synsets are connected to the same language-independent concept, such as the English synset of *'book'* and the Italian synset of *'libro'* in Fig. 3. Other relations (e.g., hypernymy, meronymy) can also be deduced across languages from concept relations (e.g., subsumption, part-of) of the interlingua.

The interlingua can also be used to provide abstraction from language-specific lexical meanings and to incorporate language-independent knowledge not provided by any underlying wordnet, for example, through the introduction of new relations among concepts. This can be reused as additional background knowledge during the ontology matching task.

Finally, the *domain layer* annotates interlingua concepts with explicit domain information. Depending on the field of study (e.g., computational linguistics, knowledge representation, information retrieval), multiple formalisations exist for the notion of domain: lexical (as categories of words), semantic (as categories of concepts), or pragmatic (as categories of documents) [13]. We adopt the semantic approach and *define a domain as a labelled category of concepts*. Within the bounds of this general definition, different concrete models may be proposed and do exist (which we will shortly present); on this abstract level of definition, however, we do not impose any other constraints on how domains should be implemented as NuSM can be adapted to any of those models.

In our implementation of the background knowledge, as lexical databases we used wordnets from the *Open Multilingual WordNet* project.[1] Our interlingua layer is an extended and modified version of the Princeton WordNet synset graph. Our domain layer, finally, is a converted, language-independent version of the *Extended WordNet Domains* [14] resource. Section 8.1 gives more details on our implementation as well as a list of alternative, freely available components from which the three-layered background knowledge of NuSM can be built.

## 5. Language aware label parsing

The goal of the first step of semantic matching (cf. Section 2) is to formalise natural language ontology labels into propositional description logic formulas. This formal representation acts as a pivot and abstracts away various aspects of heterogeneity common in natural language such as ambiguity, variations in phrasing, and also multilingualism. In NuSM, label parsing is carried out in three substeps: language detection, computation of the structure of the formula, and computation of the atoms of the formula.

[1] http://compling.hss.ntu.edu.sg/omw/.

Please cite this article in press as: G. Bella, et al., Language and domain aware lightweight ontology matching, Web Semantics: Science, Services and Agents on the World Wide Web (2017), http://dx.doi.org/10.1016/j.websem.2017.03.003.

### 5.1. Language detection

The initial step of language detection provides the language of each input tree as parameters to the matcher that, in turn, instantiates the appropriate language resources and pipelines. Language detection is performed using a state-of-the-art statistical tool [15]. We do not handle the rare case of ontologies mixing labels in multiple languages, as it is a relatively rare phenomenon and language detection on the level of individual labels tends to be inaccurate due to their shortness. Processing is interrupted if for the detected language no lexical or NLP resources are available.

### 5.2. Computing the formula structure

Label formulas are built from concepts, boolean operators, and brackets. This phase consists of NLP operations with the following goals:

- distinguish words that will become atoms (concepts) from words with connective functions (e.g., *and*, *or*) that will become logical operators;
- map each word of the latter category to its corresponding logical operator;
- compute bracketing from the syntactic structure of the label.

Traditionally, NLP is conceived as a pipeline of text annotation operations. NuSM executes a pipeline consisting of well-known NLP tasks: *tokenisation*, *part-of-speech tagging*, *lemmatisation*, *multiword detection*, and *syntactic parsing*. These are followed by the two matching-specific steps of *operator mapping* and *formula building*. We demonstrate the process on our running example of *'Growing of plants and animals'* taken from Fig. 1. The result is shown in Fig. 4.

*Tokenisation* identifies the boundaries of words and punctuation and models text as a series of tokens (the resulting tokens are marked in Fig. 4 as boxes).

*Part-of-speech tagging* identifies linguistic functions of words, which we use to distinguish between, on the one hand, *open-class words* with associated meanings (in our case: nouns, verbs, adjectives, and adverbs) that we treat as atoms in the logical formula to be built and, on the other hand, *closed-class words* (in our case: pronouns, prepositions, conjunctions, punctuation, etc.) that will later either be eliminated or become logical operators (part-of-speech tags are represented by a 'POS' annotation in Fig. 4).

*Lemmatisation* finds the canonical forms of open-class words (as they appear in our multilingual knowledge base), e.g., through morphological analysis (represented by an 'L' (lemma) annotation in Fig. 4).

*Multiword detection* finds multi-word expressions with meanings distinct from those of its component tokens, e.g., 'hot dog' (our running example does not contain multiwords).

*Syntactic parsing* extracts the phrase structure from the label, in the form of a parse tree, as shown in Fig. 5.

*Operator mapping* maps closed-class words and punctuation found between two open-class words to the logical operators of conjunction, disjunction, negation, or $\varnothing$ (nothing) according to the rules illustrated in Fig. 6 (represented by 'OP' (operator) annotation in Fig. 4).

*Formula building* takes the output of syntactic parsing and operator mapping to build a bracketed logical formula, as shown in Fig. 5 where the formula $growing \sqcap (plant \sqcup animal)$ is obtained.

In language aware label parsing, we are confronted with two specific problems with respect to conventional NLP tasks: firstly, that labels in multiple languages need to be processed and, secondly, that labels are short (between 2 and 6 tokens in our evaluation corpora) and follow a non-standard syntax. With state-of-the-art NLP tools multilingual support involves a costly process of building dedicated NLP pipelines for each language. The label shortness problem, in addition, results in poor performance from existing general-purpose NLP tools, which are typically customised for longer pieces of text such as web pages, newswire, and Wikipedia articles. Indeed, state-of-the-art tokenisers, part-of-speech taggers, parsers, etc., are implemented as machine-learning-based tools that examine a window of preceding and following words around the word being annotated. The shortness of text means that fewer contextual words are available. Furthermore, syntax and punctuation in labels are different from conventional text for which NLP resources are trained and tuned:

- verbs and adverbs are rare (verbs tend to appear as gerunds, e.g., *growing*, or participles, e.g., *grown*);
- syntactic units are most often limited to noun, adjective, and prepositional phrases;
- word order, capitalisation, and punctuation are used in non-standard ways (e.g., instead of *'growing of plants'* one may find *'Plants, Growing of'*).

As we also claim in [9], this kind of text, typically appearing in structured data, can be described as a specific type of *block language*. We argue that the block language of structured data needs a unique approach to NLP, both for reasons of efficiency in supporting new languages and due to the shortness and the specific syntactic and orthographic rules. NuSM thus uses a custom NLP framework of multilingual pipelines. First described in [9] and still under active development, the framework is specifically built for the semantic analysis of block language (i.e., its scope of usage is not limited to semantic matching). While the framework does reuse freely available NLP resources (OpenNLP models for tokenisation and POS tagging, lemma and multiword dictionaries), its pipelines and certain components are specifically tuned to the task of parsing short labels:

- a rule-based tokeniser extension was developed for non-standard orthography frequent in ontology labels;
- the output of conventional POS taggers is not entirely relied upon (because of their lower performance on short text) and is rather used as a hint in further processing steps such as lemmatisation and word sense disambiguation;
- some machine learning components were retrained on short labels for some languages, e.g., English and Italian POS taggers and named entity recognisers[2]; we are planning to retrain other components as future work on the NLP framework.

### 5.3. Computing atomic concepts

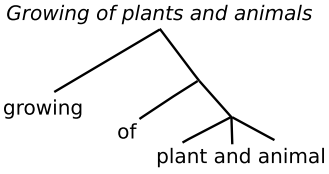
This substep of label parsing retrieves meanings for open-class words appearing as atoms in the label formula. For each word, all possible meanings are retrieved, from all possible domains. A crucial difference with respect to SMATCH and similar monolingual matchers is that the meanings retrieved are represented as language-independent concepts from the interlingua instead of English-specific WordNet synsets. Thus, for the word *'plant'* in *'Growing of plants and animals'* we retrieve from the multilingual knowledge base both the concept *'#45 plant as organism'* and the concept *'#12 industrial plant'*, reaching them via the word senses within the English lexical database (Fig. 3).

[2] Note that named entity recognition and disambiguation components are not currently used in NuSM.

Please cite this article in press as: G. Bella, et al., Language and domain aware lightweight ontology matching, Web Semantics: Science, Services and Agents on the World Wide Web (2017), http://dx.doi.org/10.1016/j.websem.2017.03.003.

| *Growing* | *of* | *plants* | *and* | *animals* |
|---|---|---|---|---|
| POS = noun | POS = closed-class | POS = noun | POS = closed-class | POS = noun |
| L = growing | OP = ⊓ | L = plant | OP = ⊔ | L = animal |

**Fig. 4.** Result of tokenisation (boxes), part-of-speech tagging ('POS'), lemmatisation ('L'), and operator mapping ('OP') on the label '*Growing of plants and animals*'.




| *Growing* | *of* | ( | *plants* | *and* | *animals* | ) |
|---|---|---|---|---|---|---|
| POS = noun | POS = closed-class | | POS = noun | POS = closed-class | POS = noun | |
| L = growing | OP = ⊓ | | L = plant | OP = ⊔ | L = animal | |

**Fig. 5.** Result of syntactic parsing and subsequent formula building on the label '*Growing of plants and animals*': bracketing is introduced.

| English | Italian | Spanish | Operator |
|---|---|---|---|
| except, non, without, ... | eccetto, escluso, non, senza, ... | excepto, sin, no, fuera de, salvo, ... | ¬ |
| and, or, ',', ... | e, ed, o, ',', ... | y, e, o, u, ',', ... | ⊔ |
| all other closed-class words | | | ⊓ |
| absence of closed-class word between two open-class words | | | ⊓ |

**Fig. 6.** Mapping of closed-class words in labels to description logic operators (the list is incomplete). For example, the phrase '*English literature except poetry*' will be mapped to *english* ⊓ *literature* ⊓ ¬*poetry*.

Furthermore, we account for phenomena such as paraphrasing and approximate translations that we observe to be more frequent across languages than in monolingual matching. As an illustration let us take the frequent case of English noun–noun compounds such as *'school transport'* or *'river tourism'*, both taken from our EUROVOC evaluation corpus. In Latin languages (Italian, Spanish, French, etc.) these are typically translated as noun+adjective: *'trasporto scolastico'*, *'turismo fluviale'*. In order to increase recall for such matches, we also retrieve meanings of *derivationally related words*, e.g., *'river–fluvial'*, *'school–scholastic–schooling–education'*, or *'cycling–bicycle'* (cf. the example lexical gap on Fig. 3). This is why we are able to match *'river tourism'* with *'turismo fluviale'* and also *'cycling tourism'* with *'turismo a bicicletta'* in the example in Fig. 1.

## 6. Domain aware label disambiguation

The retrieval of all possible meanings for each atom results in highly ambiguous labels. Finding the most relevant concepts (meanings) for polysemous words is the role of the well-known NLP task of *word sense disambiguation* (WSD).

In semantic matching, sense disambiguation serves the purpose of avoiding erroneous mappings that decrease precision and may also decrease recall (if a wrong mapping is retained instead of the right one). More precisely, it eliminates word meanings that are incorrect in the given context: e.g., an equivalence relation between *'stock'* as in a business context and *'broth'* (or *'brodo'* in Italian) in the context of cooking.

In NuSM a domain-based disambiguation method is used for two reasons: firstly, we observed that most real-world matching scenarios involve domain-specific ontologies. Secondly, our domain-based approach is independent of text length and is therefore well-suited to short ontology labels. The method relies on domain–concept mappings retrieved from the domain layer of our background knowledge (cf. Section 4) and is implemented based on our previous results from [9]. We refer the reader to this paper for more details as well as for evidence on the efficiency of domain-based sense disambiguation on the block language of structured data.

An important feature of the method is that it is entirely language-independent since performed on the level of interlingua concepts. Costly language-specific WSD solutions are thus not necessary. The process is divided into two main steps:

1. *domain detection:* the relevant domain(s) of the ontology are made explicit through automated estimation;
2. *domain-based disambiguation:* the word meanings most relevant to the detected domain(s) are selected.

Just like for language detection, the domain detector assumes the input tree to be homogeneous with respect to its domain(s). This assumption is analogous to the *one-domain-per-discourse hypothesis* in computational linguistics that claims that *'multiple uses of a word in a coherent portion of text tend to share the same domain'* [13, p. 28]. While the majority of real-world use cases does seem to fit our assumption, multidomain ontologies do exist, such as large national and international classifications in industry (NAICS[3]), commerce (SITC[4]), or libraries (UDC[5]). These, however, are invariably *faceted classifications* [16] that very clearly divide their contents domain-wise by top-level nodes. It is therefore straightforward to extract the domain-specific portions of such classifications as a preprocessing step and feed only homogeneous trees to the matcher.

[3] The North American Industry Classification System. https://www.census.gov/eos/www/naics/.

[4] Standard International Trade Classification. https://unstats.un.org/unsd/trade/sitcrev4.htm.

[5] Universal Decimal Classification. http://www.udcc.org.

Please cite this article in press as: G. Bella, et al., Language and domain aware lightweight ontology matching, Web Semantics: Science, Services and Agents on the World Wide Web (2017), http://dx.doi.org/10.1016/j.websem.2017.03.003.

The domain detector receives as input all concepts from the input classification tree, retrieved in a previous phase. Domain detection is performed on the tree as a whole: as described in [9], the relevant domains of each concept are retrieved from the domain layer (cf. Fig. 3) and, based on the distributions of domains of all concepts in all formulas in the tree, a global relevance score is computed for each domain. The domains with the highest relevance scores are the likeliest to describe the ontology. For example, in the English tree in Fig. 1 the subtree consisting of the single label *'Growing of plants and animals'* may see the domain categories of *agriculture*, *botany*, *zoology*, or *industry* associated to it due to the various meanings of its words and their typical contexts of use. Of these, *agriculture* is likely to be the most relevant domain as all three words *growing*, *plant*, and *animal* have meanings pertaining to it, while *industry* is the least relevant domain as it only pertains to the *industrial plant* meaning of the word *plant*.

All domain scores obtained above are taken as input by the domain-based disambiguator that, again using the domain–concept mappings, for each atom chooses the concept or concepts that are most relevant with respect to the domains detected. The method is parametrisable to keep not just one but the *best k* meanings, which amounts to *filtering* the least relevant senses. Setting $k = 0$ deactives disambiguation completely. In the example of *'Growing of plants and animals'*, even though the word *plant* has two meanings (according to Fig. 3), the domain of *agriculture* helps in disambiguating the correct meaning of *'#45 plant as organism'*.

## 7. Cross-lingual label matching

The role of step 3 in semantic matching is to gather axioms about the concepts appearing in the source and target trees, allowing the inference of mapping relations beyond what can be derived from mere concept or string equality. In monolingual matching, binary relations are retrieved from lexical databases (without any need to use the interlingua layer) and are used to derive propositional logic operators of implication and negation (Fig. 7, first column). In the absence of such semantic relations and also for words entirely missing from the lexical database (out-of-vocabulary words), in an attempt to increase recall, equivalence relations are derived using string-level techniques common in ontology matchers: Levenshtein-distance-based string similarity (the threshold set to 0.8) and synset gloss matching are used.

For cross-lingual matching the following changes apply:

- language-independent relations are used;
- similarity and derivational relatedness relations are used to extend lexical coverage;
- string-similarity-based matching is adapted to the cross-lingual use case;
- gloss-based matching is deactivated.

NuSM retrieves language-independent ontological relations from the interlingua. Depending on how the interlingua is built, these relations may differ from the synset relations present in wordnets and may provide additional or different knowledge. For example, concept relations in the UKC were modified to respect ontology modelling principles as opposed to psycholinguistic principles used in Princeton WordNet [12]. In the case of BabelNet, a large number of additional relations were extracted from various sources such as Wikipedia [17]. Fig. 7 shows how both language-specific lexical and language-independent ontological relations are converted by NuSM into propositional logic operators of implication and negation. The propositional formulas built this way are then used in step 4 by the SAT solver carrying out the actual node matching task.

Matching recall is improved by exploiting *similarity* and *derivational relatedness* relations whenever provided by lexical databases (these extra related concepts were retrieved in step 1.2). For *'river'* in the English label *'river tourism'* the related concept *'fluvial'* is retrieved, equivalent to the concept within the Italian label *'turismo fluviale'*.

Regarding string-based matching, our experiments have shown that decreasing the similarity threshold of word-level fuzzy string matching results in a significantly improved (5%–20%) cross-lingual matching recall while causing a slighter decrease (0%–2%) in precision. Concretely, a Levenshtein distance metric was used with the similarity threshold decreased from 0.8 (monolingual SMATCH) to 0.6 (cross-lingual NuSM). We explain this phenomenon by the different nature of the matching of out-of-vocabulary words (missing from the interlingua) between the English monolingual and the cross-lingual scenarios. In the former, the forms of matching words are either identical, carry a very slight morphological (*book–books*) or orthographic (*Montreal–Montréal*) variance, or are radically different (synonyms: *book–volume*). This is why a conservative fuzzy string matching approach (threshold $= 0.8$) is appropriate. In the cross-lingual case, between certain language pairs the orthographic proximity of words sharing a meaning is a common phenomenon: for example, *'tourism'* and *'turismo'* (similarity of 0.71). Generally, the closer the lexicon, morphology, and orthography of two languages, the more effective string matching can be between them. Accordingly, we observed the greatest increase in F-measure with the very closely related Spanish–Italian language pair (+18%) and a still fairly good increase for English–Spanish (+7%) and English–Italian (+3%) which all share a large vocabulary of Latin or French origin. Applying the same lower threshold to English monolingual matching, on the other hand, actually resulted in a significant drop of precision and F-measure, confirming the hypothesis that for English a more conservative string matching works better. For the monolingual matching of morphologically richer languages, however, lower thresholds may in turn be more appropriate in dealing with variations in affixes. Finally, in the case of languages using different writing systems (e.g., Latin and Cyrillic, Latin and Chinese), string similarity methods do not work, unless combined with more sophisticated transliteration methods.

Finally, gloss-based matching is discarded in the cross-lingual scenario as glosses tend to be absent in freely available lexical databases and, when present, are written in different languages that gloss-based matchers cannot handle.

## 8. Extensibility by languages and domain terminologies

A wide range of language and domain resources are available on the web that we can reuse in our approach. In this section we first provide an overview of existing knowledge resources from which the three-layered background knowledge, presented in Section 4, can be constructed. Then we describe the processes by which the background knowledge can be extended, considering two major scenarios:

- providing support for a new language;
- extending lexical coverage for an already supported language by generic or domain-specific terms.

The first case requires plugging a new wordnet into the background knowledge as well as providing language-specific NLP components. The second case supposes that a wordnet-structured lexicon is already integrated into the background knowledge which is being extended by the contents of a domain wordnet or term base. All of these cases are depicted in Fig. 2, p. 5.

Please cite this article in press as: G. Bella, et al., Language and domain aware lightweight ontology matching, Web Semantics: Science, Services and Agents on the World Wide Web (2017), http://dx.doi.org/10.1016/j.websem.2017.03.003.

| Lexical relations | Ontological relations | Mapped to |
|---|---|---|
| synonymy, similarity, derivationally related forms, pertainymy | identical concepts | $\leftrightarrow$ |
| hypernymy, hyponymy, holonymy, meronymy | is-a, attribute-value, part-whole, substance, membership | $\leftarrow$ or $\rightarrow$ |
| antonymy | | $\neg$ |

**Fig. 7.** Mapping of lexical relations (first column) and language-independent ontological relations (second column) to propositional logic operators used by the SAT-based matcher. Antonymy is used exclusively for monolingual matching. For hierarchical relations (such as *is-a* or *part-of*) transitivity is taken into account.

### 8.1. Reuse of existing multilingual knowledge bases

The three-layered lexical-semantic architecture we defined in Section 4 is approximated, to various extents, by several already existing and often freely available resources. The main differences among them reside in the details of their formal models and in the way they are populated (through manual effort, algorithmically, semi-automatically).

In earlier efforts such as EuroWordNet [18], MultiWordNet [19], or the Multilingual Central Repository [20], language-specific wordnets are constructed semi-automatically. Cross-lingual interoperability is provided by mapping non-English synsets to their English Princeton WordNet (PWN) counterparts. In other words, the English synset graph itself serves as the interlingua. This means that most of these multilingual resources inherit both the lexical coverage and the Anglo-Saxon lexical-semantic bias of PWN [21,22]. Nevertheless, they were and still are enormously popular and have often served as a basis for further efforts in building multilingual lexical-semantic resources.

A more recent and more extensive project is *BabelNet* [17]. It is automatically built from existing wordnets, including the *Open Multilingual WordNet* collection, as well as from other resources such as *Wikipedia*, *Wiktionary*, or *OmegaWiki*. Currently it covers 271 languages. Its interlingua layer, composed of *Babel synsets*, was enriched with a large number of synsets and relations, and is thus different in terms of organisation and richness from earlier resources.

Another recent project, developed at the University of Trento independently and in parallel with BabelNet, is the *Universal Knowledge Core* [16]. The UKC is a multi-layered knowledge resource with its lexical and interlingua layers corresponding to those described above. The current UKC supports 38 languages. Like BabelNet, it was mostly populated from existing freely available wordnets, principally from *Open Multilingual WordNet*. However, it also includes manually added linguistic knowledge, from single lexical entries to entire wordnets (e.g., the Mongolian wordnet [23]). The relative importance of manual extension and curation in the objective of maintaining a high-quality resource is a distinguishing feature of the UKC with respect to parallel solutions.

For our research we used the UKC as the underlying knowledge resource of NuSM, due to its off-line availability as well as practical reasons of pre-existing integration with our systems. However, as BabelNet and the UKC share the same architecture (as depicted in Fig. 3), the former could also be plugged in and used with NuSM, with the solutions explained in this paper remaining applicable and relevant. The modular architecture of NuSM makes the development of a BabelNet connector a relatively easy task and one to be investigated in the future.

The domain layer, as defined above, is also instantiated in several existing resources. The first such resource was *WordNet Domains* [24] where domains are labelled with strings and are defined as sets of Princeton WordNet synsets. *Extended WordNet Domains* [14] builds on the former but defines domains as fuzzy sets, i.e., domain–concept relations are annotated with weights. Finally, WordNet itself defines *domain term categories* that are represented not as labels but as synsets themselves but that, however, only categorise a small subset of the WordNet synsets.

For NuSM we implemented the domain layer as a modified version of *Extended WordNet Domains* where domains are mapped to concepts of the UKC interlingua as opposed to Princeton WordNet synsets, as described in detail in [9].

### 8.2. Extension by new wordnets

A large number of language-specific wordnets are downloadable from the web, in most cases under one of the common free licences.[6] These resources, even though not always encoded in the same file formats, tend to follow the logical structure of PWN and so can be reused for our purposes. Often developed through research or community efforts, these resources offer variable levels of lexical coverage. If a wordnet does not exist at all for a language or its coverage is deemed inadequate, various automated [25] and manual (expert-sourced or crowd-sourced [26]) enrichment methods are applicable. While the presentation of these methods extends beyond the scope of this paper, we mention as an illustration that the *Open Multilingual Wordnet* project so far managed to integrate 34 wordnets, while its extended version, generated through automated methods from Wiktionary data, integrates 150 languages [25].

The mapping of the synsets of the new wordnet to interlingua concepts is straightforward to automate for all wordnets that are synset-aligned with PWN, which is generally the case. The interconnection of language-specific wordnets is still an actively researched topic and is one of the major tasks undertaken by the *Global WordNet Association*.[7]

In the case of the UKC, as explained in Section 8.1, the interlingua is a modified version of the PWN synset graph, and a mapping resource between the two graphs is constantly maintained. For a new language – say, Spanish – we perform the mapping as shown in the pseudocode below. Note that the same method can also be applied to multilingual lexical databases other than the UKC as long as they share the architecture described above.

**Algorithm 1** Adding a new wordnet (sub)graph

```
1:  procedure addNewWordNet(newWordNet)
2:      for newSynset in traverseBFS(newWordNet) do
3:          pwnSynset ← mapToPWN(newSynset)
4:          if pwnSynset = None then
5:              concept ← attachSynset(newSynset)
6:          else
7:              concept ← mapToConcept(pwnSynset)
8:          connect(newSynset, ukcConcept)
9:  function attachSynset(synset)
10:     newConcept ← createConcept()
11:     for parent in getHypernyms(synset) do
12:         pwnParent ← mapToPWN(parent)
13:         conceptParent ← mapToConcept(pwnParent)
14:         addISARelation(newConcept, conceptParent)
15:         if domain ← getDomain(synset) = None then
16:             domain ← getDomain(conceptParent)
17:         mapToDomain(newConcept, domain)
18:     return newConcept
```

[6] E.g., from the Global WordNet Association or from Open Multilingual Wordnet.

[7] http://globalwordnet.org.

1. Through breadth-first traversal of each synset in the Spanish wordnet (line 2), retrieve the corresponding PWN synset (line 3).
2. If such a mapping exists, go to step 4 below.
3. Otherwise, first the new synset will need to be attached to a newly created concept (line 5).
   (a) Create a new concept in the interlingua (line 10).
   (b) Retrieve the parent(s) of the new Spanish synset in the input Spanish wordnet (line 11).
   (c) For each parent, retrieve the corresponding PWN parent synset (line 12) and interlingua parent concept (line 13).
   (d) Create IS-A relations between the new concept and the parent concept(s) (line 14).
   (e) As the domain category of the new concept, use the domain of the Spanish synset if the Spanish wordnet provides such information; otherwise, use the domain of the parent concept.
4. For the PWN synset obtained, retrieve its corresponding interlingua concept (MAPTOCONCEPT, line 7).
5. Connect the new synset to the interlingua concept (line 8).

Note that all functions and procedures called within the pseudocode above are basic getters or setters with the exception of the mapping functions MAPTOPWN and MAPTOCONCEPT. While the former (mappings of local synsets to Princeton WordNet synsets) is provided by each input wordnet, the latter (mappings of Princeton WordNet synsets to interlingua concepts) is maintained as part of the UKC.

### 8.3. Extension by domain terminologies

Domain terminologies also exist in wordnet format [3,27–29] that easily integrate into multilingual lexical databases such as those described above. The *Diversicon project*[8] provides a few such domain-specific lexical databases, such as the *UK Civil Protection lexicon* pertaining to the emergency response domain and presented in [3].

However, it is more common in the terminology community to encode terminology in *terminological databases* (*term bases* in short). Term bases are usually multilingual by design: they are organised around *concepts* (also known as *terminological entries*) to which one or more terms are attached per language [30, pp. 879–883]. This means that interlingual mappings are by default provided by the term base. It is straightforward to convert a term base into a new wordnet structure: a term becomes a word (lemma), its terminological entry (concept) becomes both one synset in each language and one interlingua concept, and the sense connecting the synset and the word is automatically created. Hypernymy relations are also often provided by term bases and can be reused to construct the graphs of new wordnet synsets and interlingua concepts.

Connecting such converted domain terminologies to the interlingua, however, is not a trivial process as the mappings between terminological entry concepts and interlingua concepts are typically not provided. If the term base consists of a single tree of terms or a small amount of such trees that attach to the leaves of the interlingua (highly specialised terms tend to appear at the bottom of wordnet hierarchies) then plugging it into the wordnets and into the interlingua is a straightforward operation that can be carried out manually by a knowledge engineer. If the mapping between domain and interlingua concepts is more complex (because it requires the mapping of overlapping concepts or changes in the interlingua concept graph structure) then additional knowledge engineering methodologies may be needed, such as those provided in [31,32].

[8] http://www.diversicon-kb.eu.

### 8.4. Extension by NLP resources

The languages currently supported by the NLP framework of NuSM are English, German, Spanish, Italian, Chinese, and Mongolian, while Arabic is under development.

The framework was built with scalable extensibility in mind where most of the linguistic operations are generalised either to be language-independent or shareable across languages with similar properties. The most important aspect of generalisation is that the semantic NLP tasks of meaning retrieval, formula building, and word sense disambiguation were conceived and implemented as entirely language-independent components that use the interlingua and domain layers of the background knowledge. Secondly, processing logic was separated from linguistic resources (machine learning models, lemmatisation and multiword dictionaries, operator mappings as illustrated in Fig. 6) and is reused across languages. Thirdly, whenever language-specific resources are not available, the framework is able to fall back on less precise but still effective generic solutions (e.g., Unicode-based tokenisation for lots of writing systems, syntactic parsing based on POS tags alone).

Consequently, the following is the minimal set of NLP components that need to be specifically provided for each new language to be supported:

- a *tokenizer* for languages where default Unicode-based tokenisation is not applicable (e.g., Chinese);
- a *mapping of closed-class words to operators*, as illustrated in Fig. 6, which typically contains less than 100 words;
- a *lemmatiser* that can be a simple *lemmatisation dictionary* for languages that are morphologically not too complex;
- a *syntactic (constituency) parser* where it is sufficient to cover the limited syntax of the block language of structured data.

While for best performance the part-of-speech tagger and syntactic parser components need to be adapted to block language, they can be approximated by components built for more conventional corpora.

## 9. Combining NuSM with machine translation

NuSM and translation-based matchers use radically different multilingual label parsing techniques. As evoked in the introduction, each approach has its specific strengths and weaknesses as well as its particular implementation and underlying resources. This consideration leads us to the intuitive idea of *combined matchers*: we set out to investigate the extent to which the combination of machine translation and native cross-lingual matching can improve matching results with respect to either method alone. The smaller the proportion of shared mistakes, the higher the improvement we can expect from a combined matcher.

### 9.1. A qualitative comparison of matching errors

Based on our evaluation results (as reported in Section 10) we conducted a qualitative analysis of false positive and false negative matches returned by either of the two techniques, the results of which are synthesised below.

In the case of NuSM we found the following three main reasons to be behind the large majority of mistakes:

- incompleteness of background knowledge;
- alternate wordings;
- the limitations of NLP.

Please cite this article in press as: G. Bella, et al., Language and domain aware lightweight ontology matching, Web Semantics: Science, Services and Agents on the World Wide Web (2017), http://dx.doi.org/10.1016/j.websem.2017.03.003.

By *incompleteness of background knowledge* we refer to missing words, meanings, or semantic relations. Wordnets are domain-independent resources that lack specialised domain terminology. As already discussed in Section 4, the Spanish and Italian resources we used for our evaluations have about one-third or one-fourth the lexical coverage of Princeton WordNet. We quantified the effect of lexical incompleteness on matching scores in Section 10.3.

By *alternate wordings* we refer to the capacity of language to express the same (or very similar) meaning in several ways. It may seem trivial to point out that natural language phrases and expressions often cannot correctly be translated word-by-word from one language to another. In the context of multilingual classification labels, often written in a more controlled and semi-formal language due to their normative use, the hypothesis is not as trivial but still holds according to our observations. An example is the English label *'Manufacturing'* translated into Italian as *'Attività manifatturiere'*, i.e., *'Manufacturing activities'*. The label formula computation method in SMATCH and NuSM cannot deal with approximate matches resulting from the presence of additional concepts (here: the concept of activity). Google Translate, on the other hand, is built to be able to perform phrase-level statistical translation and thus tends to be more robust in such cases. In contrast, the rule-based Apertium lacks a phrase-level statistical translation capability, leading to significantly lower matching scores.

Finally, by *limitations of NLP* we cover a wide range of linguistic processing errors often due to the inherently difficult task of parsing short text labels. Mainstream NLP tools such as machine learning models for part-of-speech tagging, parsing, etc., tend to be trained on longer conventional text such as newswire or Wikipedia articles and, hence, tend to underperform on short ontology labels. While NLP in NuSM was tuned to short labels, we observed that among NLP-related mistakes by far the most frequently recurring ones were committed by the syntactic parser, resulting in incorrect bracketing in label formulas. These mistakes can partly be attributed to the weakness of our parsing logic, partly to the inherent ambiguity of short labels (e.g., the label *'floor and wall covering'* could be correctly parsed both as $(floor \sqcup wall) \sqcap covering$ and as $floor \sqcup (wall \sqcap covering)$, the former of which representing best the intended meaning).

We found the following to be the most typical mistakes made by Google translate on our evaluation corpora.

- alternate wordings;
- committing on wrong word meanings;
- training anomalies;
- syntactic parsing mistakes;
- cumulative mistakes in non-English language pairs.

While to a different extent, *alternate wordings* are also a problem for statistical machine translation. While Google Translate is able to provide phrase-level translations which makes it inherently more robust to the phenomenon of alternate wordings than NuSM that operates on the word- and multiword-level, it nevertheless has its limitations and cannot translate, e.g., *'Psicopatologia'* (meaning *psychopathology*) into *'Abnormal psychology'*, its English equivalent in the UDC corpus.

*Committing on wrong word meanings* is caused by the machine translator needing to commit on the meaning of a polysemous input word in order to produce a single piece of translated text as output. For example, the Italian label *'Lavoro e fatica'*, literally meaning *'Work and fatigue'*, is translated by Google into *'Work and effort'*, since *'effort'* is indeed one of the meanings of *'fatica'*. The two English words *'effort'* and *'fatigue'* having distinct meanings, translation-based matching fails. NuSM is not concerned by this problem as it does not try to disambiguate meanings before matching happens: both meanings of *'fatica'* are attempted to be matched.

*Training anomalies* are due to the fact that state-of-the-art statistical machine translation is to a large extent based on sentence- or word-aligned parallel corpora used as training material, often obtained through automated processing. Errors in the original content or in the preprocessing algorithms lead to strange translation mistakes such as *'(psicotecnica)'* (appearing within parentheses in the UDC corpus and meaning *psychotechnics*) being translated into *'(Psycho)'* (meaning *psychopath*), or *'politica dell'informazione'* (meaning *information policy*) into *'political information'*.

By *syntactic parsing mistakes* we refer to choosing the wrong parsing, especially when the phrase structure is ambiguous, e.g., *'Concetti e leggi generali'* (meaning *General concepts and laws*) is translated into *'Concepts and general laws'*. This results in an incorrect label formula being built by the matcher tool.

*Cumulative mistakes in non-English language pairs* refer to a phenomenon of an increased number of matching errors when none of the two input languages is English. When matching a Spanish tree against an Italian tree, both need to be translated into English, resulting in a higher probability of translation errors appearing compared to the case where one of the trees is in English.[9] NuSM does not suffer from this effect and can even take advantage of the linguistic proximity of languages for improved results. This can be observed in our evaluations (those without OOV words, in order to eliminate the bias due to lower lexical coverage) where NuSM generally obtained better scores for the Spanish–Italian language pair (both being Romance languages sharing a similar morphology and syntax) while GoogleSM performed relatively worse.

### 9.2. Combination methods

Several approaches to combining matching techniques are possible; here we introduce two possibilities:

- as a simple fusion post-processor on mappings;
- different matchers on different inputs.

In simple fusion the two matchers are used as black boxes and only their output mappings are combined. This is the architecture depicted in Fig. 8. This solution is the simplest both conceptually and implementation-wise, but is also more limited in the kinds of combinations it supports.

Discrimination of matchers based on the label allows for a more efficient matching of deep single-domain or multidomain classifications. In such classifications, the deeper in the hierarchy labels are found the more domain-specific they tend to be. Since the most popular statistical machine translators offer good results on domain-agnostic or moderately domain-specific texts, a possible combination technique is to run translation-based matching on labels close to the root, and NuSM with its background knowledge extended by domain terminology applied on deeper levels.

While the second method appears to be a promising research direction, we have so far only experimented with the first one. The combined matcher architecture we considered is shown in Fig. 8. The output of both SMATCH and NuSM are mappings in the form of (*source-node*, *target-node*, *rel*) where *rel* is a semantic relation out of $\mathcal{R} = \{\equiv, \sqsubset, \sqsupset, \varnothing\}$. The *combiner* component is a function

$$f_c : \mathcal{R} \times \mathcal{R} \to \mathcal{R}$$

[9] Let us note that the problem would still remain if direct translation were applied from one language to another (e.g., from Spanish to Italian), as Google Translate always uses English as a pivot, resulting in two subsequent translations being performed.

Please cite this article in press as: G. Bella, et al., Language and domain aware lightweight ontology matching, Web Semantics: Science, Services and Agents on the World Wide Web (2017), http://dx.doi.org/10.1016/j.websem.2017.03.003.

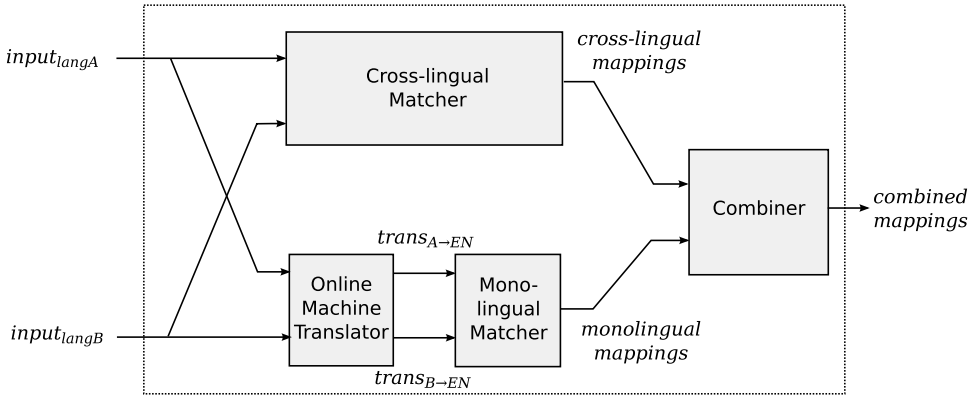



**Fig. 8.** A combined matcher that generates its output using the mappings output by its two component matchers.

| $f_\vee$ | $\equiv$ | $\sqsubset$ | $\sqsupset$ | $\varnothing$ |
|---|---|---|---|---|
| $\equiv$ | $\equiv$ | $\equiv$ | $\equiv$ | $\equiv$ |
| $\sqsubset$ | $\equiv$ | $\sqsubset$ | $\varnothing$ | $\sqsubset$ |
| $\sqsupset$ | $\equiv$ | $\varnothing$ | $\sqsupset$ | $\sqsupset$ |
| $\varnothing$ | $\equiv$ | $\sqsubset$ | $\sqsupset$ | $\varnothing$ |

| $f_\wedge$ | $\equiv$ | $\sqsubset$ | $\sqsupset$ | $\varnothing$ |
|---|---|---|---|---|
| $\equiv$ | $\equiv$ | $\sqsubset$ | $\sqsupset$ | $\varnothing$ |
| $\sqsubset$ | $\sqsubset$ | $\sqsubset$ | $\varnothing$ | $\varnothing$ |
| $\sqsupset$ | $\sqsupset$ | $\varnothing$ | $\sqsupset$ | $\varnothing$ |
| $\varnothing$ | $\varnothing$ | $\varnothing$ | $\varnothing$ | $\varnothing$ |

**Fig. 9.** Definition of combining functions $f_\vee$ (left) and $f_\wedge$ (right) based on the mapping relations output by NuSM and by the machine translation-based SMATCH. In case the two inputs do not match, $f_\vee$ returns the stronger of the two while $f_\wedge$ returns the weaker one. In case of equally strong but contradictory input mappings (e.g., $\sqsubset$ and $\sqsupset$) both functions behave in a conservative manner and output $\varnothing$ (no relation).

taking for each source–target node pair the output relations of both matchers as input and returning a new relation $rel_c$:

$rel_c := f_c(rel_{\text{NuSM}}, rel_{\text{MT}})$.

We considered two possible $f_c$ combining functions:

- a 'greedy' or 'OR-like' function $f_\vee$ that emits the stronger of the two input mapping relations;
- a 'conservative' or 'AND-like' function $f_\wedge$ that emits the weaker of the two input mapping relations.

The precise definitions of both functions are given in the tables in Fig. 9. Note that in the case of semantic matching neither SMATCH nor NuSM associates confidence scores to its mappings so the combiner can only rely on mapping relations.[10]

In a real-world use case the choice of combining function will be determined by practical needs: if priority is given to reducing the number of false positives then $f_\wedge$ should be used as it is designed to be 'conservative' and return fewer false positives, thereby increasing precision at the cost of lower recall. On the other hand, $f_\vee$ should be used in order to increase recall at the price of decreased precision, as it is defined so as to decrease the number of false negatives. As in real-world use cases, as also reflected by our evaluation results, precision tends to be much higher than recall, hence optimising by $f_\vee$ generally leads to an improved F-measure and can thus be used as an all-purpose combiner.

## 10. Evaluation and discussion

The objective of our evaluations was not only to provide precision and recall figures on NuSM itself but mainly to compare the performance of our method to that of state-of-the-art machine-translation-based monolingual matching. We did not include highly domain-specific ontologies in our evaluation corpus, nor did we extend NuSM with specialised domain terminology, in order to avoid an unfair comparison to general purpose machine translation tools that are not extensible in the same manner. In other words, to a certain extent, our evaluation results downplay one of the strengths of our approach, that is, the incremental adaptability of the background knowledge to the matching task.

We set up four separate evaluation scenarios: (1) comparison of NuSM to machine-translation-based matching over our full evaluation corpora; (2) the same comparison but over corpora guaranteed not to contain out-of-vocabulary words (i.e., not covered by our background knowledge), in order to get an idea of the effect of lexical incompleteness (or the lack thereof) on matching results; (3) evaluation of the fusion matcher; (4) evaluation of the relevance of word sense disambiguation for ontology matching.

### 10.1. Evaluation method

Our evaluations were performed on three language pairs: English–Spanish, English–Italian, and Spanish–Italian. As multilingual evaluation corpora we used the Universal Decimal Classification (UDC) and a randomly chosen subset of the EUROVOC vocabulary,[11] both available in several languages. Statistics on our evaluation corpora can be found in Fig. 10.

The two corpora we used are significantly different – and thus complement each other well – in terms of label size and complexity: EUROVOC labels tend to be very short (average length of 2.3 tokens) while UDC labels are longer (5.3 tokens). EUROVOC thus presents a use case where labels are syntactically very simple, yet their meanings are harder to identify due to reduced context. UDC, on the other hand, provides richer labels both syntactically and contextually, while the longer label presents a greater challenge for parsing with a proportionally higher probability of mistakes.

Earlier evaluations we had performed (such as those in our previous work [33]) showed that deep nesting of nodes in input classification hierarchies causes a significant drop in matching recall and thus introduces a bias into the results that makes any difference between label parsing methods appear as less pronounced. In order to focus on evaluating label matching and to avoid any interference in our results stemming from structure-level matching, we decided to 'flatten' the tree of the UDC corpus into a list of top-level nodes. EUROVOC is already a flat list of terms so it did not need to be transformed in any way. Thus, the – otherwise unchanged – step 2

[10] For other kinds of matchers that only output equivalence mappings with confidence scores, $f$ would need to compute a fusion score from the two input confidence scores.

[11] EUROVOC: the EU's multilingual thesaurus (eurovoc.europa.eu), UDC: Universal Decimal Classification (udcc.org). Of the latter we used Main Tables 1–7, excluding table 5 because it largely consists of Latin botanical and zoological named entities.

| Corpus | Languages | Nb. of nodes per language | Avg. English label length | Avg. Italian label length | Avg. Spanish label length |
|---|---|---|---|---|---|
| EUROVOC | EN, ES, IT | 589 | 1.9 | 2.2 | 2.3 |
| UDC | EN, ES, IT | 965 | 5.6 | 6.3 | 6.2 |

**Fig. 10.** Corpora used for evaluation.

of semantic matching (as described in Section 2.2) does not have any effect in our evaluations.

As these classifications are node-aligned across languages, we took these aligned nodes as ground truth for equivalence mappings. However, semantically valid subsumption and sometimes even equivalence mappings are also found to hold between unaligned nodes. Manual production of ground truth being beyond our means for the roughly 1,600 nodes of our corpora, we have simplified our evaluations in order to allow their automation:

- only relations of equivalence, that is, only perfect matches are evaluated as positives while subsumptions and disjointness are ignored;
- all returned equivalences that are not in the ground truth, even if deemed valid by a human observer, are considered as false positives.[12]

These very conservative evaluation criteria obviously affect our precision and recall scores in a negative way. We are not worried about this as the primary goal of our evaluations – at least for the purposes of this paper – is not the assessment of NuSM performance in absolute terms but rather its comparison to the approach based on machine translation, used by most other state-of-the-art multilingual matchers. Furthermore, it is important to point out that our goal was to compare label matching methods rather than complete ontology matcher tools: we wished to find out how the use of a local multilingual knowledge base (as presented in Section 4) fares against machine translation, *all other things being equal*. For this reason, we reused the original monolingual SMATCH for our evaluations that we fed with machine-translated English input labels.

The API-based Google Translate service is typically exploited in one of the two following manners:

- *sentence-to-sentence*, i.e., on the level of entire labels, obtaining a single output translation: this is the method used by most state-of-the-art matchers such as AML [4];
- *word-to-word* providing all possible translations of the input word (as one would by consulting a dictionary).

The second method was already thoroughly discussed and evaluated recently in related work [34]. For our evaluations we preferred to choose the first translation method: not only is it more frequently used in matchers but it also more markedly differs from our approach, being capable of offering translations over word *n-grams*. This provides an approximate phrase-level matching capability, something our technique cannot do or only in a limited way (for multiword expressions present in our lexicons). Our comparisons thus provide further insight on the effect of such phrase-level translations on matching results with respect to our approach that uses the Interlingua on the lexeme level.

[12] For example, on the EUROVOC corpus the matcher returned the mapping *'tube ≡ metropolitana'* which is correct if the sense of *'metro'* is assumed for *'tube'*. This mapping, however, was classified as a false positive as EUROVOC assumed the sense of *pipe* and therefore asserted a mapping to *'tubo'*.

### 10.2. Results on full evaluation corpora

Our results are shown in the left-hand column of Fig. 11, with precision in blue at the top, recall in yellow in the middle, and F-score in red at the bottom. The dark bars correspond to NuSM (natively cross-lingual matching) while the light bars represent GoogleSM (scores obtained by English-only SMATCH when fed by Google-translated English text). The occasional third bar shows results when using Apertium, the well-known rule-based machine translator [35], in lieu of Google. The version of Apertium we used for our evaluations does not support Italian, which is why it only covers the English–Spanish language pair. The rationale of including Apertium as a second machine translation service is that, contrary to Google Translate, it is a free (as in open source) tool that can be used both on-line and off-line.

Precision scores are in the 98%–99% range for the EUROVOC corpus (Apertium scoring a slightly lower 97%). They are somewhat lower (91%–92%) for UDC, which we explain by the specific syntax and punctuation used by the UDC corpus that tends to induce more false positive matches. On average, precision results are similar among NuSM, GoogleSM, and Apertium. We observe a slight (about 1%) advantage of GoogleSM that we attribute to the more aggressive fuzzy string matching we used in NuSM, as explained in Section 7, which resulted in a few more false positives.

Precision scores are generally much higher than recall, a phenomenon common for most ontology matchers (see, for example, the results of the 2014 Ontology Alignment Evaluation Initiative, [36]). In the case of recall, the overall higher scores for the EUROVOC corpus are explained by shorter labels (two words per label on average instead of six for UDC), the probability of parsing and matching mistakes being proportional to label length. On EUROVOC, recall scores for Google-based matching are by 12%–13% higher for the English–Spanish and English–Italian language pairs, and only by 3% for Italian–Spanish. Similarly, on UDC English–Spanish the difference is 12%, while on English–Italian only 4% and on Spanish–Italian our method wins by 1%. Apertium scores lowest, about 15% below our results.

We attribute the better results of GoogleSM with respect to NuSM to factors that we discuss in detail in Section 9.1. It also has to be noted that the three languages we evaluated – English, Spanish, and Italian – are among the languages with the largest available online text corpora, which makes Google Translate particularly efficient when translating from one to another. As future work, we are planning to apply our evaluations to languages with considerably lower online presence, where we expect worse performance from Google Translate.

Another insightful result is the inverse behaviour of NuSM and GoogleSM on the Spanish–Italian language pair: here, the translation-based method obtains the weakest results while NuSM is the strongest. This is explained by two main factors: firstly, performing two translations to English instead of one increases the probability of mistakes. Secondly, NuSM takes advantage of the orthographic, lexical, and syntactic similarity of Spanish and Italian: fuzzy string matching handles out-of-vocabulary words well across these languages, and their very similar syntax results in similar bracketing in label formulas. These useful cases of proximity disappear when both languages are translated to English.

Finally, the huge performance drop resulting from the replacement of Google Translate by Apertium shows that translation quality has a great effect on matching results in the SMATCH and NuSM

Please cite this article in press as: G. Bella, et al., Language and domain aware lightweight ontology matching, Web Semantics: Science, Services and Agents on the World Wide Web (2017), http://dx.doi.org/10.1016/j.websem.2017.03.003.

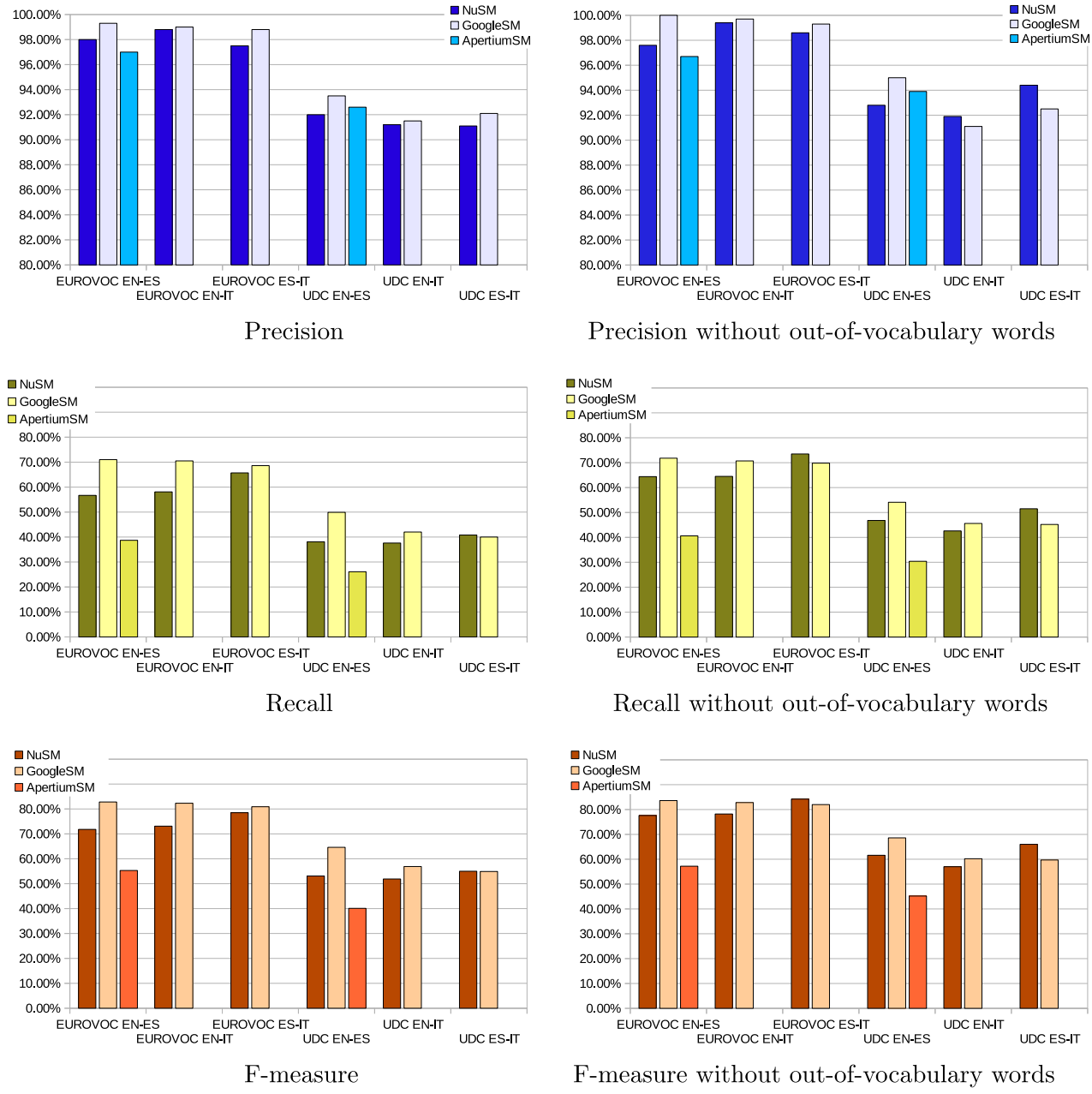



**Fig. 11.** Cross-lingual evaluation results on parallel classifications. Three language pairs (English–Italian, English–Spanish, Spanish–Italian) are considered over two data sets (EUROVOC and UDC). Dark bars: NuSM (cross-lingual matcher). Light bars: GoogleSM (English-only SMATCH using Google Translate). The third bar for the English–Spanish language pair: ApertiumSM (English-only SMATCH using the Apertium machine translator). Top (blue) row: precision, middle (yellow) row: recall, bottom (red) row: F-measure. Left-hand column: results on the full evaluation corpora, right-hand column: adjusted results on corpora not containing out-of-vocabulary words. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

matchers. The lower phrase-level accuracy of Apertium leads to matching results that are clearly outperformed by the cross-lingual method used by NuSM.

### *10.3. Results on OOV-adjusted corpora*

Out-of-vocabulary words are frequently encountered during matching if the lexical (general or domain-specific) coverage provided by the multilingual lexical database is inadequate. In our evaluations this was the case of both the Spanish and Italian wordnets, the former containing 35K synsets and 32K words and the latter 34K synsets and 40K words (the English one contained 110K synsets and 134K words). On the other hand, string similarity is a remarkably efficient method for matching OOV words across our three evaluation languages, as we showed in Section 7, and is capable of mitigating a lower lexical coverage.

In order to evaluate more precisely the effect of OOV words on our results we performed a second round of evaluations. We started by filtering out all nodes containing OOV words in order to obtain *adjusted corpora* that are free from the OOV effect. 18 to 50% of the nodes were thus filtered out from each data set pair, showing that OOV words are indeed a major phenomenon in our corpora. Note that OOV words are merely one of the possible ways a lexical database may be incomplete: missing word meanings or missing relations may also lead to false negatives or false positives. We did not consider these other cases in our evaluation.

The adjusted results on the filtered corpora are shown in the right-hand column of Fig. 11. While precision did not change

Please cite this article in press as: G. Bella, et al., Language and domain aware lightweight ontology matching, Web Semantics: Science, Services and Agents on the World Wide Web (2017), http://dx.doi.org/10.1016/j.websem.2017.03.003.

considerably, recall has improved by 5%–8% for NuSM, with a somewhat more pronounced gain on the Spanish–Italian language pair (8% on EUROVOC and 11% on UDC). On corpora without OOV words the F-score of NuSM arrives within 3%–7% of GoogleSM on English–Spanish and English–Italian, and outperforms it by 2%–6% on Spanish–Italian. The Apertium-based matcher scores much lower than the two other methods.

These numbers show that string matching, while helping a lot, cannot fully solve the problem of OOV words. Having appropriate lexical coverage in the underlying knowledge bases still results in a major improvement of matching performance.

We finally note a slight improvement both for GoogleSM and ApertiumSM on the adjusted corpora, showing that the lexical coverage of the two machine translator systems is not perfect either.

### 10.4. Results of combined matching

Evaluation results of both combination methods are shown in Fig. 12. For each data set the first two bars recap previous results for NuSM and GoogleSM, the third bar shows the result when combining by $f_\wedge$ (AND), and the fourth bar when combining by $f_\vee$ (OR). Note that, as our evaluation is solely based on equivalence relations being returned, the only combinations in Fig. 9 that affect our results are those where either of the two input relations is an equivalence.

As hypothesised, the two label parsing methods – in the native languages on the one hand, through machine translation to English on the other hand – have provided mappings that only partially overlap. This leads to the combiners consistently producing significantly better scores than either of its two input matchers. As expected, $f_\wedge$ improves precision while $f_\vee$ improves recall as well as F-measure in all test cases.

With respect to the better-performing individual matcher, by using $f_\wedge$ we manage to improve precision for EUROVOC by 0.3–0.9% (where figures were already around 98%–99%) and for UDC by 2%–5%, reaching 94%–97%. By using $f_\vee$ we boost recall figures by 7%–13% for EUROVOC and by 10%–13% for UDC. F-measure is increased in all cases by 4%–12% when using $f_\vee$. The best improvements are again reached on the Spanish–Italian language pair which benefits the most from the complementarity of the two label matching approaches.

Fig. 13 shows some additional statistics on the effect of combined matching on the mappings. Table (a) shows the total number of mappings (all types confounded) produced by NuSM and GoogleSM on each evaluation corpus. Tables (b) and (c) show the number of times $f_\wedge$ and $f_\vee$ modified a mapping, respectively.

### 10.5. Results on the relevance of word sense disambiguation

As part of the analysis of our evaluation results we assessed the impact of word sense disambiguation. We were motivated by what we saw as important mitigating factors with respect to the negative effects of polysemy on matching results.

Firstly, the accuracy of the disambiguation method is crucial as erroneous disambiguations lead to lower matching recall because relevant meanings are pruned. Secondly, identifying the correct meanings of words is, in itself, neither a sufficient nor a necessary condition of eliminating false positive mappings. For example, the label pair *'(business) stock'* and *'brodo'* actually need to appear in the ontologies being matched for a false positive mapping to materialise. Since in real-world use cases ontologies tend to be domain-specific, the likelihood of a simultaneous appearance of business- and cooking-related terminology is rather low. Thirdly, even if polysemy arises on the level of individual words, context words are likely to mitigate the problem: the full label *'company stocks'* will not be matched with *'broths for cooking'* because no relation exists between *'company'* and *'cooking'*. The longer the labels, the less likely are polysemy-induced false positives to appear.

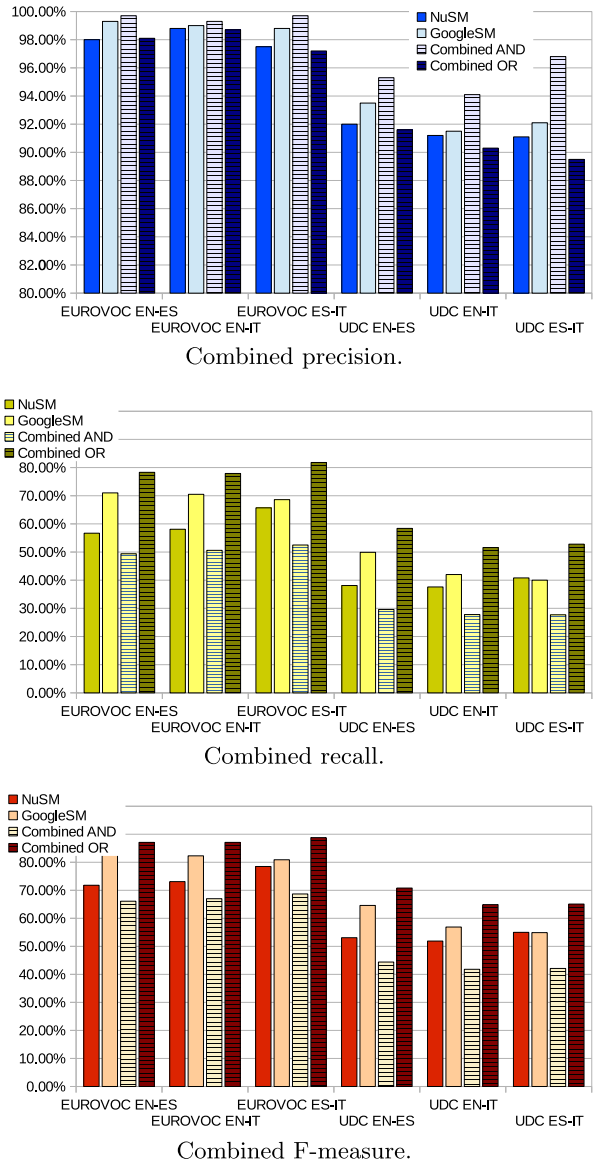



**Fig. 12.** Evaluation of the combining matchers based on $f_\wedge$ and $f_\vee$. For each data set, first bar: NuSM, second bar: GoogleSM, third bar: $f_\wedge$ (precision-optimised), fourth bar: $f_\vee$ (recall-optimised).

In order to quantify the effect of polysemy on results in NuSM, we tested the effect of the absence of disambiguation on our evaluation results (presented in Section 10). Two ambiguous atoms were considered to match if out of all possible concept combinations there was at least one pair that was found to be related. As polysemy results in the appearance of false positive matches, we manually identified false positives that were due to wrong concepts of polysemous atoms (words) being matched. Only 10% of all false positives were found to be due to the lack of sense disambiguation (the rest were due to knowledge incompleteness or to NLP-related errors such as syntactic parsing mistakes). The overall effect on precision and recall, due to the otherwise very high precision (typically in the 91%–100% range, cf. Section 10), was thus lower than 1% for our evaluation sets.

The general lesson that we take from these results is that the sense disambiguation component should primarily be applied in

Please cite this article in press as: G. Bella, et al., Language and domain aware lightweight ontology matching, Web Semantics: Science, Services and Agents on the World Wide Web (2017), http://dx.doi.org/10.1016/j.websem.2017.03.003.

| Matcher | EUROVOC | UDC |
|---|---|---|
| NuSM | 6,435 | 25,750 |
| GoogleSM | 6,220 | 23,054 |

(a) Total number of mappings created by each matcher on each corpus.

| $f_\wedge$ | EUROVOC | UDC |
|---|---|---|
| $\equiv \mapsto \sqsubset, \sqsupset$ | 133 | 404 |
| $\equiv \mapsto \varnothing$ | 395 | 496 |
| $\sqsubset, \sqsupset \mapsto \varnothing$ | 2,138 | 15,504 |

(b) Number of mappings modified by $f_\wedge$.

| $f_\vee$ | EUROVOC | UDC |
|---|---|---|
| $\sqsubset, \sqsupset \mapsto \equiv$ | 133 | 404 |
| $\varnothing \mapsto \equiv$ | 395 | 496 |
| $\varnothing \mapsto \sqsubset, \sqsupset$ | 2,125 | 15,431 |

(c) Number of mappings modified by $f_\vee$.

**Fig. 13.** Statistics on the influence of combining functions on mappings.

scenarios of low precision, such as very short and highly ambiguous labels.

## 11. Related work

A domain aware technique is used for ontology matching in [37]. Here, the matching configuration is highly asymmetric as the target ontology is DBpedia with a huge amount of nodes. The objective is thus to filter the contents of DBpedia to entities similar to the input. Furthermore, domains are not predefined with respect to lexical meanings but are computed with respect to the matching task in an unsupervised and ad-hoc manner from structural information taken from DBpedia, such as instance–class or superclass relations. The ontology labels themselves do not play any role in the computation of domains and the technique is entirely language-agnostic. In conclusion, this method is usable in combination with an existing matcher tool in scenarios where the target ontology is large and provides ample structural context for robust domain definitions. Our approach, in contrast, also works on small inputs in symmetric matching scenarios and assumes that the semantics of nodes are in a large part contained within the labels as opposed to structure and relations. This is the case of lightweight ontologies that are the main target of NuSM.

The dominant approach to cross-lingual ontology matching is to translate ontology labels to a common target language, thereby reducing the problem to monolingual (most often English-to-English) matching. State-of-the-art matchers rely either on Bing or the Google Translate API, including AML [4] and LogMap [6], the two tools that performed best in the *Multifarm* cross-lingual matching tasks of the 2014 Ontology Alignment Evaluation Initiative [36], the most authoritative evaluation effort for ontology matchers. Likewise, oft-cited publications on multilingual and cross-lingual matching [5,7,8] all propose methods that rely on some form of translation, using either online services or dictionaries either on the level of whole labels or on the level of individual words.

The idea of using interconnected lexical databases for cross-lingual matching also appears in the recent paper [34], in the context of a comparison between BabelNet and Google Translate as online word-level translation services. Beyond this initial similarity, our approach is conceptually different. The matching technique described in [34] uses BabelNet and Google Translate not as multilingual knowledge bases but, again, as online translator services that retrieve all possible *translations of words* (i.e., lemmas) appearing in labels. A simple form of meaning-level reasoning is introduced by telling the online service to augment the set of returned lemmas by synonyms. Matching is thus performed on the word level of a chosen pivot language. We, on the other hand, perform matching directly on the language-independent level of concepts where beyond synonymy we are able to exploit a richer set of concept relations such as *subsumption* or *part-of*.

The ontology label matching problem can be reformulated as that of *textual similarity* or *entailment*: if $a \rightarrow b$ (textual entailment) then $a \sqsubseteq b$ (subsumption label mapping). The approach taken by NuSM can thus be regarded as a knowledge- and logic-based cross-lingual textual entailment operation (with the added complexity of formalising labels in the context of their ancestor nodes). Cross-lingual textual entailment has received some interest in the last years. The backbone of state-of-the-art solutions is usually a statistical approach (e.g., machine learning on parallel corpora [38], cross-lingual distributional semantics [39]) or – in most cases – a simple machine translation to a pivot language (English) [39,40]. They are sometimes combined with a knowledge-based approach for handling cases of monolingual synonymy and polysemy [40]. These methods are not fundamentally different from what we evaluated by combining Google Translate with English-to-English semantic matching, and present similar strengths and weaknesses: on the one hand, they can achieve good performance when the underlying machine translator, word embedding, or machine learning model is of high quality and is appropriate to the matching task. On the other hand, high quality is reached through access to large parallel or comparable training corpora, while appropriateness to the matching task requires the same corpora to be close enough to the domains to which the input belongs. Another particularity of the statistical approaches to textual similarity or entailment is their tendency to gloss over small differences that fundamentally change the meaning of a phrase. For example, the two phrases *'cereals and rice'* and *'cereals except rice'*, the likes of which often appear in classifications, tend to be found very similar by statistical methods while they will be properly handled by NuSM that is able correctly to convert the two phrases into strictly non-matching formulas.

## 12. Conclusions

We have presented a new approach to semantic ontology matching that uses natively language and domain aware techniques, relying on off-line multilingual NLP and lexical-semantic resources. The results we obtained confirmed the viability of the method. When compared to the state-of-the-art cross-lingual matching technique based using two different machine translation tools, the three approaches turned out to score roughly similarly in terms of precision. In terms of recall, Google Translate reached equivalent to slightly better scores (+0%–15% depending on the language pair) while the Apertium machine translator fared much worse (−15%–20%).

We found our slightly lower scores with respect to Google to be partially due to the incompleteness of our local multilingual resources (both concerning lexical coverage and NLP processes). Indeed, with complete lexical coverage the differences in recall between the two methods are greatly reduced and our method takes the lead by 2%–6% on the Spanish–Italian language pair. This is a significant observation considering the fact that in the case of our matcher the issue of lexical incompleteness is under the control of the user: coverage issues can be – and are expected to be – addressed through the enrichment of background lexical-semantic knowledge by domain terminology and facts.

Finally, we built a fusion matcher that exploits the differences between the knowledge- and translation-based approaches by

Please cite this article in press as: G. Bella, et al., Language and domain aware lightweight ontology matching, Web Semantics: Science, Services and Agents on the World Wide Web (2017), http://dx.doi.org/10.1016/j.websem.2017.03.003.

combining their resulting mappings. The combining functions we defined can be used for increasing either precision or recall. We were able consistently and significantly to increase the combined system's precision or recall in comparison to the better of its two components. We consider these as positive and insightful results, especially in light of the simplicity of the fusion methods.

We explain the success of the fusion matcher by the inherent and irreducible differences between the knowledge- and the translation-based method. The former is based on finding correspondences between the meanings of lexemes across languages, and is not adapted to dealing with paraphrasing and approximations across languages (unless those phrases are lexicalised). State-of-the-art statistical machine translation, operating on the word *n-gram* level, is more robust with respect to this problem. Yet, it has its own limitations: its essentially corpus-driven approach leads to much weaker support for languages and domains more sparsely represented on the web. Furthermore, being a translation tool, it needs to commit on specific natural-language translations and is thus prone to disambiguation errors on short labels.

Our evaluations also confirmed the intuitive hypothesis that the quality of the machine translation service largely influences the results of the cross-lingual matcher built on top of it, much like the quality of our lexical and NLP resources influences NuSM. In the case of statistical machine translation, quality is dependent on the size of the parallel or comparable corpora available for each language pair. Another condition is the semantic closeness of these corpora to the application domain: for example, machine translators trained on multilingual legal corpora issued by the European Parliament can hardly be used to translate, say, medical texts heavily laden with domain terms. In order to reach a good level of recall, both translation-based and knowledge-based matcher systems (supposing they both are under the user's full control) need to be extended for matching tasks that either cover a yet unsupported language or domain. For machine translators, the state of the art uses significant amounts of appropriate parallel or comparable corpora, while for knowledge-based matchers knowledge acquisition methods (manual or automated) and a core set of language-specific NLP components are necessary. In the end, the preference for either method may depend on the availability of the respective resources.

## Acknowledgements

This work was supported by the *SmartSociety* project, funded by the European Commission's 7th Framework Programme (grant number 600854).

## References

[1] Fausto Giunchiglia, Maurizio Marchese, Ilya Zaihrayeu, Encoding classifications into lightweight ontologies, in: Journal on Data Semantics VIII, in: Lecture Notes in Computer Science, vol. 4380, Springer, Berlin, Heidelberg, ISBN: 978-3-540-70663-2, 2007, pp. 57–81.

[2] Jérôme Euzenat, Pavel Shvaiko, Ontology Matching, Second Ed., Springer, ISBN: 978-3-642-38720-3, 2013.

[3] Francisco Quesada, Fiona McNeill, Gábor Bella, Alan Bundy, Improving dynamic information exchange in emergency response scenarios, in: Proceedings of the 14th Annual Conference for Information Systems for Crisis Response and Management, ISCRAM, 2017.

[4] Daniel Faria, et al., The AgreementMakerLight Ontology Matching System, in: Robert Meersman, et al. (Eds.), On the Move to Meaningful Internet Systems: OTM 2013 Conferences, in: Lecture Notes in Computer Science, 8185, Springer, Berlin, Heidelberg, ISBN: 978-3-642-41029-1, 2013, pp. 527–541.

[5] Bo Fu, Rob Brennan, Declan O'Sullivan, Cross-lingual ontology mapping –an investigation of the impact of machine translation, in: Asunción Gómez-Pérez, Yong Yu, Ying Ding (Eds.), The Semantic Web: Fourth Asian Conference, ASWC 2009, Shanghai, China, December 6-9, 2009. Proceedings, Springer, Berlin, Heidelberg, ISBN: 978-3-642-10871-6, 2009, pp. 1–15. http://dx.doi.org/10.1007/978-3-642-10871-6.1. http://dx.doi.org/10.1007/978-3-642-10871-6_1.

[6] Ernesto Jiménez-Ruiz, Bernardo Cuenca Grau, LogMap: Logic-based and scalable ontology matching, in: The Semantic Web ISWC 2011, vol. 7031, ISBN: 978-3-642-25072-9, 2011, pp. 273–288.

[7] Dennis Spohr, Laura Hollink, Philipp Cimiano, A machine learning approach to multilingual and cross-lingual ontology matching, in: Lora Aroyo, Chris Welty, Harith Alani, Jamie Taylor, Abraham Bernstein, Lalana Kagal, Natasha Noy, Eva Blomqvist (Eds.), The Semantic Web –ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I, Springer, Berlin, Heidelberg, ISBN: 978-3-642-25073-6, 2011, pp. 665–680. http://dx.doi.org/10.1007/978-3-642-25073-6.42.

[8] Cássia Trojahn Dos Santos, Paulo Quaresma, Renata Vieira, An API for multilingual ontology matching, in: Proc. 7th conference on Language Resources and Evaluation Conference, LREC, Valletta, Malta, 2010, pp. 3830–3835. https://hal.inria.fr/hal-00793285.

[9] Gábor Bella, Alessio Zamboni, Fausto Giunchiglia, Domain-based sense disambiguation in multilingual structured data, The Diversity Workshop at the European Conference on Artificial Intelligence, ECAI, 2016.

[10] Fausto Giunchiglia, Aliaksandr Autayeu, Juan Pane, S-Match: An open source framework for matching lightweight ontologies, Semant. Web (ISSN: 1570-0844) 3 (3) (2012) 307–317. http://dx.doi.org/10.3233/SW-2011-0036.

[11] Fausto Giunchiglia, Mikalai Yatskevich, Pavel Shvaiko, Semantic matching: Algorithms and implementation, in: Stefano Spaccapietra, Paolo Atzeni, François Fages, Mohand-Said Hacid, Michael Kifer, John Mylopoulos, Barbara Pernici, Pavel Shvaiko, Juan Trujillo, Ilya Zaihrayeu (Eds.), J. Data Semant. 9 (2007) 1–38.

[12] George A. Miller, WordNet: A lexical database for english, Commun. ACM (ISSN: 0001-0782) 38 (11) (1995) 39–41. http://dx.doi.org/10.1145/219717.219748. http://doi.acm.org/10.1145/219717.219748.

[13] Alfio Gliozzo, Carlo Strapparava, Semantic Domains in Computational Linguistics, Springer-Verlag, Berlin, Heidelberg, ISBN: 978-3-540-68158-8, 2009.

[14] Aitor González-Agirre, German Rigau, Mauro Castillo, A graph-based method to improve wordnet domains, in: Alexander Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing: 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Proceedings, Part I, Springer, Berlin, Heidelberg, ISBN: 978-3-642-28604-9, 2012, pp. 17–28. http://dx.doi.org/10.1007/978-3-642-28604-9.2.

[15] Shuyo Nakatani, (2010) https://github.com/shuyo/language-detection.

[16] Fausto Giunchiglia, et al., Faceted lightweight ontologies, in: Conceptual Modeling: Foundations and Applications, vol. 5600, Springer, Berlin Heidelberg, ISBN: 978-3-642-02462-7, 2009.

[17] Maud Ehrmann, et al. Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC-2014, Reykjavik, Iceland, May 26-31, 2014.

[18] Piek Vossen (Ed.), EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Norwell, MA, USA, ISBN: 0-7923-5295-5, 1998.

[19] Emanuele Pianta, Luisa Bentivogli, Christian Girardi, MultiWordNet: Developing an aligned multilingual database, in: Proceedings of the First International Conference on Global WordNet, 2002, pp. 21–25, http://multiwordnet.fbk.eu/paper/MWN-India-published.pdf.

[20] J. Atserias, The MEANING multilingual central repository, in: Proceedings of the Second International WordNet Conference, 2004, pp. 80–210.

[21] Piek Vossen, Wim Peters, Julio Gonzalo, Towards a Universal Index of Meaning, in: SIGLEX99: Standardizing Lexical Resources, 1999, pp. 81–90. http://www.aclweb.org/anthology/W99-0512.

[22] Noam Ordan, Shuly Wintner, Hebrew WordNet: A test case of aligning lexical databases across languages, Int. J. Transl. 19(1) (2007) 39–58.

[23] Amarsanaa Ganbold, Feroz Farazi, Fausto Giunchiglia, An experiment in managing language diversity across cultures, in: International Conference on Information, Process, and Knowledge Management, eKNOW, Barcelona, Spain, 2014.

[24] Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, Alfio Gliozzo, Using domain information for word sense disambiguation, in: The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems, SENSEVAL '01, Stroudsburg, PA, USA, 2001, pp. 111–114. Association for Computational Linguistics, http://dl.acm.org/citation.cfm?id=2387364.2387391.

[25] Francis Bond, Ryan Foster, Linking and extending an open multilingual wordnet, in: 51st Annual Meeting of the Association for Computational Linguistics, Sofia, 2013.

[26] Fausto Giunchiglia, Mladjan Jovanovic, Mercedes Huertas-Migueláñez, Khuyagbaatar Batsuren, Crowdsourcing a large-scale multilingual lexico-semantic resource, in: The Third AAAI Conference on Human Computation and Crowdsourcing, HCOMP-15, San Diego, CA, 2015.

[27] Fausto Giunchiglia, GeoWordNet: A resource for geo-spatial applications, in: Proceedings of ESWC 2010, pp. 121–136.

[28] Luisa Bentivogli, Andrea Bocco, Emanuele Pianta, ArchiWordNet: Integrating wordnet with domain-specific knowledge, in: Proceedings of the Second Global Wordnet Conference, 2004.

[29] Paul Buitelaar, Bogdan Sacaleanu, Extending synsets with medical terms, in: Proceedings of the 1st International WordNet Conference, January 21-25, o.A., Mysore, India, 2002.

[30] S.E. Wright, G. Budin, Handbook of terminology management, Application-oriented Terminology Management, J. Benjamins, ISBN: 9789027221551, 2001. https://books.google.it/books?id=UYm7XvBXm7QC.

[31] Bernardo Magnini, Manuela Speranza, Merging global and specialized linguistic ontologies, in: Proceedings of the Workshop Ontolex-2002 Ontologies and Lexical Knowledge Bases, LREC-2002, 2002, pp. 43–48.

[32] Antonio Toral, Monica Monachini, Claudia Soria, Montse Cuadros, German Rigau, Wauter Bosma, Piek Vossen, Linking a domain thesaurus to WordNet and conversion to WordNet-LMF, in: Proceedings of Second International Conference on Global Interoperability for Language Resources, ICGL2010, Hong Kong, 2010.

[33] Gábor Bella, Fausto Giunchiglia, Ahmed Ghassan Tawfik AbuRa'ed, Fiona McNeill, A multilingual ontology matcher, in: Proceedings of OM-2015 located at ISWC 2015, CEUR-WS, vol. 1545.

[34] Mamoun Abu Helou, Matteo Palmonari, Mustafa Jarrar, Effectiveness of automatic translations for cross-lingual ontology mapping, J. Artif. Intell. Res. (JAIR) 55 (2016) 165–208.

[35] Mikel L. Forcada, et al., Apertium: a free/open-source platform for rule-based machine translation, Mach. Trans. (ISSN: 1573-0573) 25 (2) (2011) 127–144. http://dx.doi.org/10.1007/s10590-011-9090-0.

[36] Zlatan Dragisic, et al., Results of the ontology alignment evaluation initiative 2014, in: ISWC 2014, Riva del Garda, Trentino, Italy, 2014, pp. 61–104. http://ceur-ws.org/Vol-1317/oaei14_paper0.pdf.

[37] Kristian Slabbekoorn, Laura Hollink, Geert-Jan Houben, Domain-Aware ontology matching, in: International Semantic Web Conference, Springer, 2012, pp. 542–558.

[38] Yashar Mehdad, Matteo Negri, José Guilherme C. de Souza, FBK: cross-lingual textual entailment without translation, in: Proceedings of the First Joint Conference on Lexical and Computational Semantics, in: SemEval '12, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 701–705. http://dl.acm.org/citation.cfm?id=2387636.2387755.

[39] Eneko Agirre, Carmen Banea, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, Janyce Wiebe, Semeval-2016 task 1: semantic textual similarity, monolingual and cross-lingual evaluation, in: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016, pp. 497–511. http://aclweb.org/anthology/S/S16/S16-1081.pdf.

[40] Julio Castillo, Marina Cardenas, Sagan: A machine translation approach for cross-lingual textual entailment, in: Proceedings of the First Joint Conference on Lexical and Computational Semantics, in: SemEval '12, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 721–726. http://dl.acm.org/citation.cfm?id=2387636.2387759.

Please cite this article in press as: G. Bella, et al., Language and domain aware lightweight ontology matching, Web Semantics: Science, Services and Agents on the World Wide Web (2017), http://dx.doi.org/10.1016/j.websem.2017.03.003.