A multilingual ontology matcher

A Multilingual Ontology Matcher

Gábor Bella*, Fausto Giunchiglia†, Ahmed AbuRa'ed†, and Fiona McNeill*

*Heriot-Watt University, †University of Trento

Abstract State-of-the-art multilingual ontology matchers use machine translation to reduce the problem to the monolingual case. We investigate an alternative, self-contained solution based on *semantic matching* where labels are parsed by multilingual natural language processing and then matched using a language-independent knowledge base acting as an interlingua. As the method relies on the availability of domain vocabularies in the languages supported, matching and vocabulary enrichment become joint, mutually reinforcing tasks. In particular, we propose a vocabulary enrichment method that uses the matcher's output to detect and generate missing items semi-automatically. Vocabularies developed in this manner can then be reused for other domain-specific natural language understanding tasks.

1 Introduction

Classification hierarchies, tree-structured data schemas, taxonomies, and term bases are widely used around the world as simple, well-understood, semi-formal data and knowledge organisation tools. They often play a normative role both as a means for classification (of documents, open data, books, items of commerce, web pages, etc.) and as sources of shared vocabularies for actors cooperating in a given domain. Activities such as international trade and mobility rely on the interoperability and integration of such resources across languages. Cross-lingual[1] ontology matching attempts to provide a solution for creating and maintaining alignments for such use cases.

State-of-the-art matchers that evaluate as the best in the *Multifarm* cross-lingual matching tasks of OAEI [6], such as AML [1] or LogMap [9], use online translation services (typically from Microsoft or Google) in order to reduce the problem of language diversity to the well-researched problem of monolingual English-to-English matching. The success of these methods is dependent on the availability of the translation service that is being used as a black box. Still, with the constant improvement of such services, matchers using machine translation are able to provide usable results and are able to deal with a wide range of languages.

In this paper we investigate a different perspective on cross-lingual matching that considers the building and maintenance of multilingual vocabularies as part

[1] We use the term *cross-lingual matching* as a specific case of multilingual matching when ontologies in two different languages are being aligned.

of the alignment task. The method is based on the use of locally available multilingual lexical-semantic *vocabularies*. Such resources are in constant evolution and are often available on the web with a more or less wide coverage of different terminological domains.

We are motivated by three considerations: first, we set out to explore to what extent such a linguistically-oriented, non-statistical approach to cross-lingual matching can be used as a viable alternative to machine translation. Secondly, we wish to provide a natively multilingual matcher that is entirely under the control of its user and does not rely on a non-free external translator service. This is necessary for high-value applications, such as e-commerce or libraries, where quality has to remain fully under the user's control. Finally, besides using vocabularies as resources for matching, we show how the matcher's output itself can become a resource in the purpose of vocabulary enrichment. This positive feedback loop exploits mismatches for increased terminological coverage which, in turn, improves subsequent matching results. One example use case is integration of open data—available in multiple languages—for mobility applications where geographical concepts and names are matched with the *GeoWordNet* catalogue [2].

While there is existing work [7] on using post-processing to repair a matching through the enrichment of background knowledge, our goal is different: we attempt to collect missing *vocabulary elements* that can be stored and subsequently reapplied, whereas [7] finds unknown *relations* between labels that may not be reusable outside the context of the matching task.

We took as basis for our work the SMATCH semantic matcher tool, for two main reasons: first, it operates on the level of meanings of labels instead of surface techniques, which makes it a suitable tool for cross-lingual semantic comparisons. Secondly, SMATCH is designed for matching *lightweight ontologies*, semi-formal knowledge organisation structures typically used for purposes of classification, that we believe are the main focus of most real-world cross-lingual matching challenges. Lightweight ontologies, as defined in [3], are characterised by (1) having a tree structure, (2) having nodes expressed as well-formed natural language labels, (3) they assume classification semantics (the extension of a node *Italy* under a node *Literature* are documents on Italian literature), and (4) the meaning of edges is not formally defined (they may stand for *is-a*, *part-of*, etc.).

The result of this work is NuSMATCH (NuSM for short), a first step in the direction of a new-generation multilingual matcher that has built-in capabilities for cross-lingual matching and that can also be used as a multilingual vocabulary enrichment tool.

The rest of the paper is organised as follows. Section 2 presents the *multilingual knowledge base*, the core resource for our matcher. Section 3 provides a brief reminder on semantic matching and on NuSM, while section 4 details our multilingual extensions. Section 5 presents vocabulary enrichment using erroneous mappings output by the matcher. Section 6 provides evaluation results and discussion, while section 7 presents issues not yet resolved.

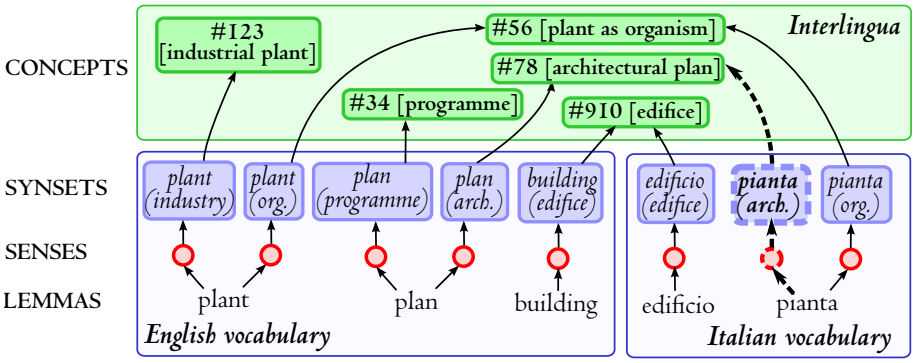

Figure 1. English and Italian vocabularies with the interlingua acting as a language-independent interoperability layer. The vocabularies may not be complete: the Italian sense and synset *pianta*, meaning 'architectural plan', is marked with dashed lines to indicate that it is missing from the Italian vocabulary.

2 A Multilingual Knowledge Base as Interlingua

Our approach to cross-lingual matching relies on a multilingual knowledge resource consisting of two layers: (1) a lower layer of multilingual *vocabularies* that are WordNet-like lexical-semantic resources; and (2) the *interlingua*: a language-independent ontology of concepts, each one linked to its corresponding vocabulary items in each language. This architecture has already been implemented at the University of Trento as part of a larger knowledge resource called the *Universal Knowledge Core* (UKC) [3], that we reuse for our purposes.

The architecture of a *vocabulary* is similar to that of Princeton WordNet [10], 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 meaning and, indeed, the WordNet synset graph is sometimes used as an upper ontology for general reasoning tasks. This practice is suboptimal because of the known Anglo-Saxon cultural and linguistic bias of the synset graph (see, for example, [12]). As a solution, our multilingual knowledge base (simply *knowledge base* in the following) introduces the *interlingua* as a manually curated ontology representing a language-independent abstraction from the synset graph. Each synset in each vocabulary is mapped to a concept (fig. 1). The opposite is not necessarily true, e.g., when a vocabulary is incomplete. The interlingua acts as an interoperability layer across language-specific vocabularies, a feature that we use for cross-lingual matching.

High-quality vocabularies are costly to build in terms of human effort. Existing wordnets[2]—that we reuse to bootstrap our vocabularies when it is legally and technically possible—tend to be incomplete to a smaller or greater extent: for

[2] `http://globalwordnet.org/wordnets-in-the-world/`

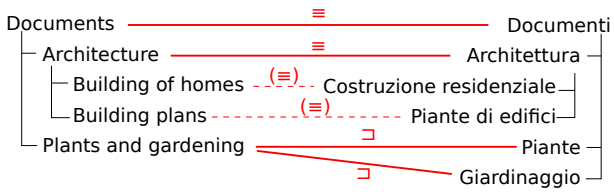

Figure 2. Example English and Italian classifications of documents, with some example mapping relations. Dashed lines with '(≡)' denote false negatives (mappings not found by the matcher), for reasons explained in section 5.

example, the Spanish *Multilingual Central Repository 3.0*[3] contains 56K lemmas and 38K synsets, the *Italian MultiWordNet*[4] contains 42K lemmas and 33K synsets, while Princeton WordNet 3.0 contains about 200K and 118K, respectively. Furthermore, wordnets tend to be general-purpose vocabularies that lack domain-specific terminology.

Efforts parallel to ours for building multilingual knowledge resources do exist. In earlier efforts such as EuroWordNet [11] or MCR [4] cross-lingual interoperability was provided by mapping non-English synsets to their English Princeton WordNet counterparts. This meant inheriting the English-centric lexical-semantic bias both in vocabulary construction and in reasoning. *BabelNet* [5] is a more recent and more advanced effort, with the same architectural design and underlying ideas as our knowledge base. The difference lies in the methodology of building it: BabelNet is mostly built automatically from diverse sources such as *Wikipedia* and *OmegaWiki*, while our knowledge base is built and maintained by human effort using both expert input and crowdsourcing. While the general problem of constructing lexical-semantic resources is beyond the scope of this paper, one of the outcomes of our work is a method for vocabulary enrichment using the output of NuSM.

3 NuSM

NuSM is designed as a multilingual extension of the SMATCH (English-only) semantic matcher [8]. Matching is semantic because, first, it is based on word senses extracted from ontology labels, secondly, it is performed using propositional logical inference and, thirdly, the mappings returned are description logic relations of equivalence, subsumption, and disjointness (for an example see fig. 2). We follow the basic four-step design of SMATCH, shown as pseudocode in fig. 3. Two new pre- and post-processing steps were added for language detection and for the semi-automated enrichment of vocabularies, respectively.

Below we provide a brief overview of each step of the matching process, followed by an in-depth discussion on the steps that are new or were modified.

[3] `http://adimen.si.ehu.es/web/MCR`
[4] `http://multiwordnet.fbk.eu`

	SMATCH	NuSM
step 0		`srcLang := detectLanguage(srcTree)` `trgLang := detectLanguage(trgTree)`
step 1	`computeLabelFormulas(srcTree)` `computeLabelFormulas(trgTree)`	`computeLabelFormulas(srcLang, srcTree)` `computeLabelFormulas(trgLang, trgTree)`
step 2	`computeNodeFormulas(srcTree)` `computeNodeFormulas(trgTree)`	

step 3	`for each srcAtom in srcTree:`
	`  for each trgAtom in trgTree:`
	`  wordNetMatcher(srcAtom, trgAtom)`
	`  stringMatcher(srcAtom, trgAtom)`
		`for each srcAtom in srcTree:`
		`  for each trgAtom in trgTree:`
		`  conceptMatcher(srcAtom, trgAtom)`
		`  nameMatcher(srcAtom, trgAtom)`

step 4	`mappings := treeMatcher(srcTree, trgTree)`	
step 5		`enrichVocabularies(mappings)`

Figure 3. Comparison of the high-level steps in SMATCH and NuSM.

For a more detailed presentation of semantic matching and the original SMATCH tool, we refer the reader to [8].

Step 0 is a new pre-processing step that detects the language of the two trees in input. We do not handle the rare case of ontologies mixing labels in multiple languages, as this would reduce the overall accuracy of language detection. Processing is interrupted if for the detected language no suitable vocabulary or NLP parser is available.

Step 1 computes *label formulas* for the two trees, that is, a propositional description logic formula corresponding to the semantic representation of the label. Atoms of the formula are sets of concepts from the interlingua, possibly representing the meaning of the atom, while operators are conjunctions, subjunctions, and negations. For example, in fig. 2, for the English label *Plants and gardening* the formula $plant \sqcup gardening$ is computed where *plant* and *gardening* are sets of concepts and the coordinating conjunction *and* becomes a disjunction (since the node classifies documents about any of the two topics). As for the label *Building plans*, it becomes a conjunctive formula: $building \sqcap plan$. The difference with respect to SMATCH is that label formulas are computed in a language-dependent manner, while meanings associated to the atoms are language-independent concepts from the interlingua instead of WordNet synsets.

Step 2 computes for each node tree their *node formulas*, which are formulas describing labels in the context of their ancestors. This step consists of computing for each label formula its conjunction with the label formulas of all of its ancestors. For *Plants and gardening*, this becomes $(plant \sqcup gardening) \sqcap document$. This step was not modified with respect to the original SMATCH.

Step 3 collects axioms relevant to the matching task. For each meaning in each atom of the source tree, step 3 retrieves all relations that hold between it and all meanings of all atoms in the target tree. In SMATCH, WordNet is used as a knowledge base (`wordNetMatcher` method) and additional axioms are inferred through string matching techniques (`stringMatcher` method). In NuSM, the interlingua is used as background knowledge (`conceptMatcher`) and string

matching is used mainly for names (`nameMatcher`). For example, for the pair of atoms (*plant*, *pianta*) retrieved from the interlingua in fig. 1, if both have a concept set of two concepts, this means retrieving potential relations for four concept pairs.

Step 4 performs the matching task (`treeMatcher` method) 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 (*overlap*) relation is returned. This step was not modified with respect to the original SMATCH.

Step 5 is introduced specifically for NuSM as a post-processing step. Its goal is to discover mismatches resulting from missing vocabulary items, and help extend the vocabulary accordingly. For example, in fig. 2, no relation is returned between *Building plans* and *Piante di edifici* if the meaning 'plan' for *pianta* is missing from the Italian vocabulary.

4 Cross-Lingual Matching

In this section we explain how steps 1 and 3 were extended to adapt to cross-lingual operation.

4.1 Computing Label Formulas

The `computeLabelFormulas` method consists of three substeps: (1) building the label formula by parsing each label using language-specific NLP techniques; (2) computing of concept sets for each atom of the label formula; and (3) context-based sense filtering for polysemy reduction.

In NuSM, word senses in label formulas are represented by language-independent concepts from the interlingua. In order to compute label formulas and the concept sets of its atoms, language-dependent parsing is performed on labels.

Substep 1.1: label formulas are built by recognising words and expressions that are to be represented as atoms, and by parsing the syntactic structure of the label. For this purpose we use NLP techniques adapted to the specific task of ontology label parsing, distinguished by the shortness of text (typically 1-10 words) and a syntax that is at the same time limited (mostly noun, adjective, and prepositional phrases) and non-standard (varying uses of punctuation and word order). Depending on the language, different NLP techniques are used:

- word boundaries are identified through language-dependent tokenisation, e.g., *dell'/acqua* in Italian vs. *water/'s* in English, the apostrophe falling on different sides;
- language-dependent part-of-speech tagging helps in distinguishing open- and closed-class words where the former (nouns, verbs, adjectives, adverbs) become atoms while the latter (coordinating conjunctions, prepositions, punctuation, etc.) become logical operators;

English	Italian	Operator
except, non, without, ...	eccetto, escluso, non, senza, ...	$\neg$
and, or, ',', ...	e, o, ',', ...	$\sqcup$
of, to, from, against, for, ...	di, del, della, dello, dell', a, al, alla, allo, all', per, contro, ...	$\sqcap$

Figure 4. Mapping of closed-class words in labels to description logic operators (the list is incomplete).

- lemmatisation (morphological analysis of word forms in order to obtain the corresponding lemmas) is also performed using language-dependent methods, e.g., rule-based, dictionary-based, or the combination of the two;
- multiwords (e.g., *hot dog*) are recognised using dictionary lookup in the appropriate knowledge base vocabulary;
- closed-class words (pronouns, prepositions, conjunctions, etc.) and certain punctuation are mapped to the logical operators of conjunction, disjunction, and negation where mappings are defined for each language (cf. fig. 4);
- syntactic parsing—that determines how logical formulas are bracketed—is also done in a language-dependent manner.

Substep 1.2: concept sets are computed for each atom by retrieving from the interlingua all possible language-independent concepts for each open-class word appearing in the label. Thus, for the word *plant* we retrieve both the concept *plant as organism* and the concept *industrial plant* (fig. 1). What is new with respect to SMATCH is the language-independence of concepts and that concepts of derivationally related words are also retrieved, e.g., *plantation*, *planting*. This provides us increased robustness with respect to approximate grammatical correspondences between labels, a phenomenon that we observed as much more common in the cross-lingual than in the monolingual case (e.g., *piante di banane* vs. *banana plantation*).

Substep 1.3: sense filtering. In SMATCH, two atoms are by default considered equal if they have the same word form or lemma, regardless of the actual meanings: if the word *plant* appears both in the source and the target tree, they may be matched regardless of their respective meanings (*living organism* or *industrial building*). In order to reduce false positives due to such cases of polysemy, SMATCH implements a form of word sense disambiguation called *sense filtering*. This operation has a lesser importance in a cross-lingual scenario as the coincidence of homographs across languages is much rarer. For example, matching the English word *plant* with the Italian word *pianta*, both polysemous as shown in fig. 1, does not pose a problem as *pianta* does not have a meaning of 'industrial plant', nor does *plant* mean 'architectural plan'. This phenomenon acts as a 'natural' word sense disambiguation technique, allowing us to finetune recall by switching off the sense filtering algorithm implemented in SMATCH when the source and target languages are different and only apply it if the two languages are the same.

4.2 Retrieval of Axioms

SMATCH performs semantic matching between atoms by retrieving axioms as WordNet relations between senses and synsets (the `wordNetMatcher` method in fig. 3). NuSM, in contrast, relies on language-independent ontological relations existing in the interlingua (`conceptMatcher`). Equivalence is implied by concept equality and subsumption is derived from *is-a*, *attribute-value*, and *part-whole* relations, taking transitivity into account.

String similarity is a common metric used in monolingual matchers. SMATCH relies on string similarity between words and between glosses of WordNet synsets (the `stringMatcher` method includes both techniques) whenever WordNet does not provide any semantic axioms. Even though string similarity has a more limited scope of use in cross-lingual matching—words unrecognised because missing from the vocabularies cannot be assumed to match across different languages—we still use it for the matching of names and acronyms which tend to have a higher resemblance across languages (`nameMatcher`). We discarded gloss-based matching as these are not available for all vocabularies and the gloss-based matcher does not work on glosses written in different languages.

5 Vocabulary Enrichment

Term lists, taxonomies, and classifications, when available in multiple languages, are useful resources for the extraction of domain-specific terminology. The idea is to exploit incorrect mappings in order to identify the vocabulary elements missing for a given language and, consequently, to enrich them in a semi-automated manner, supervised by a human user.

Generally, we consider that mappings perceived by the user as incorrect can be explained by thee main phenomena: (1) the incompleteness of the knowledge base, (2) the design and limitations of the matcher (e.g., NLP errors or the inability to match rough translations such as *Building of homes* vs. *Costruzione residenziale*, 'residential construction') , and (3) modelling errors in the classifications themselves (example: *Gardening and landscaping* classified under *Gardening* results in two being inferred to be equivalent due to classification semantics).

In the following we concentrate on errors of type 1 and especially on missing vocabulary items: word forms, lemmas, senses, and synsets. We leave the problem of enrichment of the interlingua by concepts and relations for future work. We provide a semi-automated method that identifies errors stemming from an incomplete vocabulary and proposes a corresponding repair-by-enrichment action to the user. The semi-automated approach strikes a balance between reducing human effort and maintaining the high quality of vocabularies. It requires the contribution of a skilled person, ideally a data scientist, with a good knowledge of both languages.

Step 1: selection of the tree to process. In order to detect whether vocabulary enrichment is necessary, we either rely on a decision by the user or on a heuristic based on the number of unrecognised words found in one of the trees

being over a certain threshold. The goal is to select the tree that corresponds to the vocabulary poorer in terminological coverage: in the following we will call this tree the 'poor tree' and the other one the 'rich tree'. The repair process traverses the poor tree in depth-first order from the root, as the repair of a node affects all of its descendants.

Step 2: node-by-node identification of false negative mappings. False negatives, by definition, are true mappings not found by the matcher. Our repair method, however, relies on this information to identify missing vocabulary items. For this reason, we need to have access to ground truth in the form of equivalences and subsumptions. We propose three possible methods for obtaining ground truth:

- *user-provided*, e.g., by manually pointing out false negatives node by node during the traversal process.
- *Pre-existing:* a great number of lightweight ontologies are available on the web in multiple languages, often as industry standards of economic areas englobing multiple countries (in section 6 we provide concrete examples). These multilingual classifications can be seen as *fully aligned parallel corpora* and be used for vocabulary enrichment where the alignment provides ground truth.
- *Automatically obtained:* the (monolingual) SMATCH is run in parallel using a machine translation service as preprocessor. We automate the identification of false negatives by comparing the mappings output by both SMATCH and NuSM. Negatives output by NuSM that are positives for SMATCH are likely candidates for false negatives. We assume that precision is high (false positives are few) in the monolingual case—which is generally true, cf. the evaluations in [8]—and that the overlap of the positives of SMATCH and NuSM is not total, in other words, that the former is able to provide new positives to the latter. Our experiments showed this to be the case (cf. section 6).

Step 3: identification of the missing vocabulary item and repair. As an example for the repair process, let us take the labels *Building plans* and *Piante di edifici* from fig. 2. They are represented here as atoms containing their meanings retrieved from the interlingua in fig. 1:

Building #910 'edifice'	⊓	*plans* #78 'architectural plan' #34 'programme'	≡	*Piante* #56 'plant as organism'	⊓	*edifici* #910 'edifice'

Because of the missing sense and synset 'architectural plan' for the lemma *pianta*, indicated by dashed lines in fig. 1, the equivalence is missed by the matcher. In the repair scenario, however, we are supposing it to be provided as ground truth. Once such an erroneous mapping has been identified, repair proceeds through the substeps below.

Substep 3.1: pre-selection of atoms that are likely subjects for repair. For each false negative mapping identified while traversing the poor tree, the atoms of the corresponding label are analysed. Atoms of unrecognised words (word forms or lemmas) are given priority, as an unrecognised word is a trivial cause

of false negatives. In the absence of unrecognised words, all atoms of the label are selected. In our example, the word *piante* is a recognised word (it does have one meaning, 'plant as organism', in the vocabulary), thus both $atom_{piante}$ and $atom_{edifici}$ are pre-selected.

Substep 3.2: selection of repair candidates. A repair candidate is a pair (*preselected atom*, *repair concept*) that, when the repair concept is substituted into the atom, repairs the mapping so that the mapping relation corresponds to the ground truth. In our example, ($atom_{piante}$, 'architectural plan') is such a repair candidate. In substep 2 a small subset of *repair concepts* is selected, depending on the ground truth relation to be obtained. If the relation is equivalence then the set of repair concepts corresponds to the concepts appearing in the 'rich' node formula of the mapping. If the relation is more general (resp. less general) then it corresponds to the concepts appearing in the 'rich' node formula plus all of their ancestors (resp. descendants). The suitable (*atom*, *repair concept*) pairs are retained as *repair candidates*. For the node *Piante di edifici* two repair candidates are found: ($atom_{piante}$, 'programme') and ($atom_{piante}$, 'architectural plan'). No other substitution of any concept from the left-hand side into any atom on the right-hand side leads to equivalence.

Substep 3.3: identification of the missing vocabulary item and its creation. The user filters appropriate repair candidates by answering questions such as *'is meaning "architectural plan" suitable for word* piante *in this label?'*. Upon an affirmative answer, we find the missing vocabulary item(s) within the path between the repair concept and the surface word form of the atom. Repair ends by inserting newly created item(s) into the vocabulary (again upon user acceptance). In our case, the presence of an Italian synset connected to the concept of 'architectural plan' is verified. As it is missing, a new synset is created, together with a sense and links connecting the synset with the lemma *pianta*. The created items are the ones shown in dashed lines in fig. 1.

6 Evaluation and Discussion

Our evaluations were performed on two language pairs: English-Spanish and English-Italian. We used a diverse set of industrial and public multilingual classifications and term bases.[5] As these classifications are fully aligned across languages, they provide ground truth for equivalent mappings. However, because of the nature of semantic matching, other valid equivalences and subsumptions may be returned between non-aligned nodes. For example, *Forestry/Logging* and *Forestry/Logging/Logging* are equivalent nodes according to classification semantics (both are formalised as *forestry* $\sqcap$ *logging*), yet such relations are missing from our ground truth. Manual production of ground truth being beyond our means for the 2,600 nodes evaluted, we have simplified our evaluations in order to allow the automation of tests:

[5] NACE: Statistical Classification of Economic Activities in the European Community, Rev. 2 (`ec.europa.eu/eurostat/ramon/`), EUROVOC: the EU's multilingual thesaurus (`eurovoc.europa.eu`), UDC: Universal Decimal Classification (`udcc.org`).

Corpus	Lang.	# nodes per tree	Avg. label length	Avg. depth	NuSM Prec. ≡	NuSM Recall ≡	Google smatch Prec.	Google smatch Recall
EUROVOC	EN-ES	300	2.3	1	95.9%	47.0%	98.2%	73.5%
EUROVOC	EN-IT	300	2.2	1	97.7%	56.4%	97.9%	77.9%
NACE	EN-ES	880	5.9	3.5	75.9%	20.7%	82.0%	28.5%
NACE-ATECO	EN-IT	880	6.2	3.5	82.4%	20.1%	90.3%	21.7%
UDC	EN-ES	125	5.3	2.5	63.3%	24.8%	100%	19.2%
UDC	EN-IT	125	5.1	2.5	100%	20.8%	71.7%	26.4%

Figure 5. Cross-lingual evaluation results on parallel classifications. Also included are the scores obtained by the monolingual SMATCH coupled with Google Translate.

- only relations of equivalence, that is, only perfect matches are evaluated as positives (subsumptions and disjointness are discarded);
- all returned equivalences that are not in the ground truth and cannot be trivially mapped to it (by reordering labels or removing duplicate labels) are considered as false positives.

Our results are in fig. 5. We consider the scores as promising first results, especially given our conservative evaluation method. According to close scrutiny, mapping errors (false positives and negatives) were a consequence of the following factors:

- the Spanish and Italian vocabularies we used contain 32K and 42K words, respectively, unlike our 130K English vocabulary. Missing words, senses, and synsets reduce both recall and precision.
- a weak point of our current matcher is its multilingual syntactic parser, which often results in wrong bracketing in label formulas. The longer the labels the higher the probability of a parsing error, which explains the gradual performance degradation correlated with increased label lengths in our evaluation datasets.
- the most important cause of low recall figures is the high number of non-exact translations present in the data (similar to the example *Building of homes* vs. *Costruzione residenziale*) in fig. 2). Such linguistic 'fuzziness' is perhaps the hardest cross-lingual matching problem to tackle.

The last two columns in fig. 5 represent scores obtained by SMATCH when fed by Google-translated English text. These scores are somewhat higher, although by varying margins and not in all cases. This is explained by radically different underlying NLP techniques: machine translators are essentially statistical tools based on word n-grams and thus work well on rough translations where no word-by-word cross-lingual correspondence exists. On the other hand, the statistical nature of machine translation sometimes introduces translation errors. The hypothesis that the two different approaches yield partly different matching results is confirmed by preliminary quantitative evaluations that gave 38.7% (EUROVOC), 55.3% (NACE), and 45.8% (UDC) as the percentage of true positives that were *not* found by NuSM among those that *were* found by Google-SMATCH. This proves that the translation-based method for obtaining ground truth that we supposed in section 5 can effectively work.

7 Conclusions and Future Work

The results presented in this paper, both regarding cross-lingual matching and vocabulary enrichment, reflect work in progress, with improvements ongoing in several areas. Improved language-specific syntactic parsing of ontology labels is likely to have a big impact on our scores. In the repair method, we plan to extend the scope of repair to the interlingua, both to concepts and relations. Finally, given our results, we see a new line of research in combining the vocabulary-based technique presented here with machine translation. Our observation on the difference between the sets of true positives returned by the two techniques points in the direction of a potentially efficient ensemble method.

Acknowledgment We owe a big thanks to Aliaksandr Autayeu, one of the main developers and the current maintainer of monolingual SMATCH, for his advice and for his relentless work on keeping the tool up to date. We also acknowledge the *SmartSociety* project, funded by the 7th Framework Programme of the European Community.

References

1. Daniel Faria et al. The AgreementMakerLight Ontology Matching System. In Robert Meersman et al., editor, *On the Move to Meaningful Internet Systems: OTM 2013 Conferences*, volume 8185 of *Lecture Notes in Computer Science*, pages 527–541. Springer Berlin Heidelberg, 2013.
2. Fausto Giunchiglia et al. GeoWordNet: A Resource for Geo-spatial Applications. In *Proceedings of ESWC 2010*, pages 121–136.
3. Fausto Giunchiglia et al. Faceted Lightweight Ontologies. In *Conceptual Modeling: Foundations and Applications*, volume 5600. Springer Berlin Heidelberg, 2009.
4. J. Atserias et al. The MEANING Multilingual Central Repository. In *In Proceedings of the Second International WordNet Conference*, pages 80–210, 2004.
5. 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*.
6. Zlatan Dragisic et al. Results of the Ontology Alignment Evaluation Initiative 2014. In *ISWC 2014, Riva del Garda, Trentino, Italy.*, pages 61–104, 2014.
7. Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich. Discovering Missing Background Knowledge in Ontology Matching. In *Proceedings of ECAI 2006, Riva Del Garda, Italy.*
8. Fausto Giunchiglia, Mikalai Yatskevich, and Pavel Shvaiko. Semantic Matching: Algorithms and Implementation. *J. Data Semantics*, 9:1–38, 2007.
9. Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau. LogMap: Logic-Based and Scalable Ontology Matching. In *The Semantic Web – ISWC 2011*, volume 7031, pages 273–288. 2011.
10. George A. Miller. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41, November 1995.
11. Piek Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA, 1998.
12. Piek Vossen, Wim Peters, and Julio Gonzalo. Towards a Universal Index of Meaning. In *SIGLEX99: Standardizing Lexical Resources*, pages 81–90, 1999.