What's in a 'nym'? Synonyms in Biomedical Ontology Matching

Catia Pesquita[1], Daniel Faria[1], Cosmin Stroe[2], Emanuel Santos[1], Isabel F. Cruz[2], and Francisco M. Couto[1]

[1]Dept. de Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2]Dept. of Computer Science, University of Illinois at Chicago, USA
cpesquita@di.fc.ul.pt

Abstract. To bring the Life Sciences domain closer to a Semantic Web realization it is fundamental to establish meaningful relations between biomedical ontologies. The successful application of ontology matching techniques is strongly tied to an effective exploration of the complex and diverse biomedical terminology contained in biomedical ontologies. In this paper, we present an overview of the lexical components of several biomedical ontologies and investigate how different approaches for their use can impact the performance of ontology matching techniques. We propose novel approaches for exploring the different types of synonyms encoded by the ontologies and for extending them based both on internal synonym derivation and on external ontologies.
We evaluate our approaches using AgreementMaker, a successful ontology matching platform that implements several lexical matchers, and apply them to a set of four benchmark biomedical ontology matching tasks. Our results demonstrate the impact that an adequate consideration of ontology synonyms can have on matching performance, and validate our novel approach for combining internal and external synonym sources as a competitive and in many cases improved solution for biomedical ontology matching.

Keywords: Ontology Matching, Synonym Derivation, Ontology Extension, Biomedical Ontologies

1 Introduction

Research in the Life Sciences, and in particular in biomedical research, has much to gain from Semantic Web technologies due to the amount and complexity of the data involved. One crucial development has been the creation of ontologies that describe biomedical knowledge and support several applications, both theoretical and practical, such as the representation of encyclopedic knowledge, semantic search and query, data exchange and integration, and reasoning support [1]. However, to fully benefit from the overall knowledge contained in those ontologies, meaningful connections need to be established across the concepts from various ontologies. To establish these relations, we can use ontology matching techniques that are able to find correspondences between semantically related

entities belonging to different ontologies [2].

The matching of biomedical ontologies poses considerable challenges, given their particular characteristics. The domains they cover are usually complex and large, with many biomedical ontologies possessing tens of thousands of classes dedicated to highly specific areas such as genomics, phenotypes or cellular structures. Moreover, biomedical terminology is characterized by ambiguity and complexity, features that further complicate the application of many ontology engineering techniques. However, the biomedical domain also presents some interesting opportunities such as the exploration of an abundant scientific literature or the availability of many related biomedical ontologies. Despite the efforts of the community to provide orthogonal ontologies [3], many contain overlapping knowledge. For instance, in BioPortal [4], a portal for biomedical ontologies, there are currently 306 ontologies distributed by categories, of which 59 in health, 38 in anatomy and 21 in biological processes.

In recent years the OAEI (Ontology Alignment Evaluation Initiative) [5] has been the major playfield for biomedical ontology alignment, both in its anatomy track, and more recently in the large biomedical ontologies track. An important finding of the OAEI is that many of the anatomy ontologies correspondences are rather trivial and can be found by simple string comparison techniques. To confirm this finding, a simple string matching algorithm, LOOM, was applied to several ontologies available in the NCBO BioPortal, obtaining high levels of precision in most cases [6]. Explanations for this fact include the simple structure of most biomedical ontologies, the high number of synonyms they contain, and their low language variability. Several strategies have been used by the top ranked systems at OAEI to increase recall that go beyond internal lexical similarity, including the use of external knowledge resources (SAMBO [7]) and ontologies (GOMMA [8], AgreementMaker [9]), global similarity computation techniques (AgreementMaker [10], SOBOM [11]), and more complex measures of label and structural similarity (AgreementMaker, LogMap [12]). A combination of these strategies has enabled two of the best systems, GOMMA and AgreementMaker, to reach a F-measure above 90% in the anatomy track. With the introduction of the large biomedical ontologies track in 2012, competing systems developed strategies to handle the very large size of the ontologies therein, including the selection of specific portions of the ontologies to apply matching [13]. Likewise, a new emphasis on the coherence of the generated alignments, prompted several systems to incorporate strategies to improve their alignments coherence [8, 12]. However, this shift in ontology matching systems to ensure the ontological quality of their strategies and results has not translated to the handling of terminological properties, despite the common knowledge of their importance to support matching.

The purpose of this paper is to show the positive impact that is brought by a deep understanding of the terminology contained in ontologies, when used in conjunction with current ontology matching approaches. To support this premise, we have surveyed the terminological component of several biomedical ontologies (including those used by the OAEI tracks) with a special emphasis on syn-

onyms, and tested several novel approaches to improve lexical based matching approaches. These approaches include: (1) the ranking and weighting of names and synonyms based on their degree of closeness; (2) the derivation of new synonyms based on the ones encoded by a single or both ontologies; and (3) the addition of new synonyms based on cross-references or lexical matches to related external ontologies.
The paper is organized as follows: Section 2 describes the terminological component of several biomedical ontologies and discusses their implications for ontology matching. Section 3 describes our three approaches to improve lexical-based matches. Section 4 describes the evaluation methods, while Section 5 presents and discusses the results obtained using those methods. Finally Section 6 contextualizes our contributions including their limitations and future work.

2 Synonyms in Biomedical Ontologies

Biomedical terminology is complex and ambiguous—frequently the same entity has several names (e.g., gluconeogenesis, glucose synthesis and glucose biosynthesis, all refer to the same metabolic process), a common word refers to a biomedical entity (e.g., hedgehog, and fruitfly are both gene names), or even the same word can be applied to two different entities (e.g., lingula, can either be a structure of the brain or of the lung). These challenges provide one of the major motivations to develop biomedical ontologies, given their explicit definition of concepts through ontological properties.
Biomedical ontologies characteristically have a strong terminological component in the form of names and multiple types of synonyms. Most ontologies define a primary name or label for each class, which is usually encoded as either a localname property or a label property when localnames are reserved for alphanumeric identifiers. Since biomedical entities usually have more than one name, ontologies encode alternative labels as different kinds of synonym properties, which help distinguish between the main label of a class and its alternatives, be they equivalent or merely related. Ontologies under the Open Biomedical Ontologies initiative [3] usually encode the following synonyms types: *hasExactSynonym*, where the alias exhibits true synonymy; *hasBroadSynonym* and *hasNarrowSynonym* where the aliases are broader or narrower than the primary name; and *hasRelatedSynonym*, where the alias is related to the primary class name but not necessarily broader or narrower. Other biomedical ontologies usually also encode distinct types of synonyms, reflecting different degrees of closeness in meaning to the main term. To the set of main labels and synonyms we henceforth call *names*. Some biomedical ontologies have cross-reference properties that connect ontology classes to classes from other ontologies. These links can be used to transfer name properties between cross-referenced classes. Table 1 presents some statistics on synonyms and cross-references for several biomedical ontologies, namely those provided by the OAEI, which will be used as a testbed for our proposed approaches. Most ontologies encode several synonyms for each class, with the notable exception of SNOMED, where synonyms are very rare. At the

other end of the spectrum we have UBERON, an ontology designed to integrate cross-species anatomy, which encodes a high number of distinct synonym properties, as well as cross-references to several other ontologies, including MA, NCI, SNOMED and FMA.

Although state of the art ontology matching systems use synonyms in their

Table 1: Name properties in biomedical ontologies.

Ontology	Classes	Name properties		Names per class
NCI_Human (OAEI)	3304	label	3304	1.59
		hasRelatedSynonym	5264	
MA (OAEI)	2739	label	2739	1.13
		hasRelatedSynonym	345	
FMA (OAEI)	79042	label	133629	1.69
NCI (OAEI)	66917	label	175972	2.63
SNOMED (OAEI)	122464	label	122566	1.00
FMA	78977	label	105490	1.92
		hasExactSynonym	45996	
NCI	96717	FULL_SYN*	303121	4.13
		label	96717	
UBERON	8659	label	8659	12.11
		hasExactSynonym	20955	
		hasRelatedSynonym	6150	
		hasNarrowSynonym	562	
		hasBroadSynonym	442	
		hasDbXref**	68068	

*equivalent to hasExactSynonym, **link to an external ontology or resource

strategies, they do so without considering the ontological property that encondes them and its meaning. In ontologies encoding more than one kind of name property it makes sense that ontology matching techniques differentiate between them.

3 Methods for Exploring the Use of Synonyms in Ontology Matching

3.1 Synonym Ranking and Weighting

Considering that several ontologies encode distinct types of synonyms, we base our approach on the notion that a synonym should contribute to the similarity score between two ontology classes in proportion to its closeness to the main name of the class it belongs to. To arrive at this weight, we first rank the synonyms encoded in an ontology according to the synonym property they are assigned to. Following the logical definition of commonly used synonym properties, we propose the following default ranking of name properties: (1) localname, (2)

label, (3) exact synonym, (4) related synonym, (5) broad synonym, (6) narrow synonym, (7) other synonyms. Whenever an ontology does not possess one of these properties, the rank of the following properties can be increased. This is especially relevant when matching an ontology where the localname corresponds to a unique alphanumeric identifier to an ontology where the localname is the main label of the class. These ranks can then be used to attribute weights to a classs names given the input of a single interval according to:

$$weight = 1.0 - (interval * (rank - 1)) \quad (1)$$

3.2 Ontology Lexicon Extension through Synonym Derivation

Despite the already high number of synonyms present in most biomedical ontologies, it is a cumbersome task for ontology developers to cover all possible variants. Moreover, when ontologies belong to similar but parallel domains (for instance, when they cover the anatomy of distinct mammal species) they will encode the synonyms that belong to their strict domain, but many times forgo synonyms of broader spectrum. One strategy that can be used to circumvent this omission is to extend the synsets of ontology classes with WordNet synonyms [14]. However, in the biomedical domain this strategy has been shown to slightly increase recall but at a higher cost of precision [15], which is likely due to the highly specialized vocabulary contained in biomedical ontologies and its limited coverage by WordNet.

Our novel approach is based on the notion that we can explore the synonymy relations established between sets of names within the ontologies to derive new synonyms. A preliminary implementation of this approach was integrated in AgreementMaker in 2011 [9]. The main idea behind this approach is that by finding common terms (both single and multi-word) between ontology synonyms we can infer a synonym relation between the remaining distinct terms. These terms can then be used to generate new synonym names. Since this approach is solely based on ontology terminology, we expect it to avoid the issues encountered when using a non-specific resource such as WordNet. For example, in the mouse anatomy ontology the class named as 'stomach serosa' has the synonym 'gastric serosa', which supports the inference that the terms 'stomach' and 'gastric' are synonymous. These synonymous terms are then used to create novel synonyms, by substituting terms with their synonyms in existing names. For instance, we can create a new synonym for the class 'stomach secretion' using the synonyms 'stomach' and 'gastric' to create the new synonym 'gastric secretion'.

We implement our approach in two main steps: (1) the construction of one or two thesauri containing synonym terms; and (2) the derivation of new synonyms based on thesaurus entries. The thesauri can be built based on a single ontology, one for each ontology, or based on both ontologies, resulting in a single thesaurus. This means that when new synonym terms are derived, they can be based on synonym terms inferred from the same ontology, or from both ontologies. We name these two options as intra- and inter-ontology synonym derivation, respectively. This approach is described in Algorithm 1, where creating a thesaurus T

is achieved by finding the overlapping portion of the names of each class c in an ontology O, and inferring a synonym relation between the non-overlapping portion. Extension of synonyms through derivation is based on the computation of all ontology classes' names n-grams, which can then be replaced by appropriate thesaurus entries. This approach is described in Algorithm 2.

Algorithm 1 Create thesaurus from name properties

```
input: O
  T ← ∅
  for each c ∈ O do
    names ← c.getNames()
    for each n1 ∈ names do
      for each n2 ∈ names do
        common_term ← n1.overlap(n2)
        n1_synonym_term ← n1.remove(common_term)
        n2_synonym_term ← n2.remove(common_term)
        T.add(n1_synonym_term, n2_synonym_term)
      end for
    end for
  end for
  return T
```

Algorithm 2 Extend synonyms based on derivation

```
input: O, max
  //get all n-grams of all ontology names with sizes [1,max]
  names ← O.getNames()
  ngrams ← names.getNgrams(max)
  for each ngram ∈ ngrams do
    names_n ← names.contain(ngram)
    thes_n ← T.get(ngram) //thesaurus entries that match the n-gram
    for each name ∈ names_n do
      class ← O.getClass(name)
      for each t ∈ thes_n do
        new_name ← name.replace(ngram, t)
        class.addNewName(new_name)
      end for
    end for
  end for
```

We also propose another approach to create new synonyms that is based on removing common words (i.e., words that convey little information) from the beginning or the end of names, such as 'structure' in 'spinal nerve structure'. To identify common words, we compute the evidence content for each word present in ontology names, according to the inverse logarithm of its frequency [16], then select those below a given evidence content threshold. Then for each name, we create a new synonym where leading and trailing common words are removed. We have called this approach Common Word Removal Synonym Extension (CW-SynExt), and describe it in Algorithm 3.
Coupled with this strategy, we have implemented a weighting method, where

the weight of the newly created synonym is equal to the weight of the original name multiplied by a confidence factor, which is given by the total evidence content of the synonym divided by the total evidence content of the original name. Thus, the lower the total evidence content of the removed words is, the closer the synonym captures the information conveyed by the original name and the higher will be its confidence factor.

Algorithm 3 Extend synonyms based on common word removal

```
input: O
  for each name ∈ O.getNames() do
    new_name = name
    //checks leading words
    for each word ∈ new_name do
      while word ∈ common_words do
        new_name ← new_name.remove(word)
      end while
    end for
    //checks trailing words
    for each word ∈ new_name.reversed() do
      while word ∈ common_words do
        new_name ← new_name.remove(word)
      end while
    end for
    if new_name ≠ name then
      class ← O.getClass(name)
      class.addNewName(new_name)
    end if
  end for
```

3.3 Ontology Lexicon Extension Using External Ontologies

Given the abundance of biomedical ontologies with overlapping domains, it makes sense to capitalize on correspondences to a mediating ontology to help derive the final correspondences between the ontologies to align [17]. A mediating ontology can be particularly helpful if it contains a large number of synonyms. This approach of matching a mediating ontology to each ontology and then use these results to arrive at the final alignment has been successfully used by several ontology matching systems in the biomedical domain [9, 8].

However, many biomedical ontologies encode cross-references to external ontologies, which represent relationships between classes belonging to distinct ontologies. To the best of our knowledge these have never been explicitly explored by ontology matching systems. These cross-references can be used to extend the lexicon of the ontologies being matched, by adding the name properties of the cross-referenced class to the class of the ontology being matched. For instance, the UBERON ontology encodes cross-references to the Mouse Anatomy ontology, which means that all names and synonyms of an UBERON class that references a Mouse Anatomy class can be added to its synset. This strategy bypasses the need to rely on a lexical matching between the ontologies, since the transference of the names is based on the ontology defined properties.

4 Evaluation

To evaluate our approaches that use synonyms in biomedical ontology matching we use the AgreementMakerLight system [18], a lightweight framework based on the AgreementMaker system [19], which has been optimized to handle the matching of larger ontologies. AgreementMakerLight supports a wide variety of matching methods, called *matchers*, which can be used in series or parallel such that the results from several matching algorithms can be combined into a single final result, and where correspondences are filtered by a similarity threshold. It is based on the same approaches of AgreementMaker, which have achieved top results in OAEI tracks in several years [20–22].

To remain focused on lexical approaches to ontology matching, we restrict our evaluation to two matchers that are based on the pairwise comparison of ontology classes: a name-based matcher and a word-based matcher. These matchers correspond to commonly used techniques, which are used across several other ontology matching systems (e.g., GOMMA, LogMap and YAM++ [23]). The name-based matcher (NM) consists of a straightforward comparison of the full labels or synonyms of ontology classes. The word-based matcher (WM) relies on the comparison of the words belonging to the labels or synonyms of classes through a weighted Jaccard similarity based on the evidence content of words within ontologies [18]. Although we implement our approaches as extensions to the AML framework, they are independent from it and can be used with any ontology matching system that uses lexical-based matching. To maintain further the independence of our approaches from any specific configurations of AML, we choose to combine the results of matchers through a simple join, and select them based on an empirically chosen threshold of 0.6.

We test our approaches on four matching tasks proposed by OAEI: (1) Mouse Anatomy (MA) - NCI Human Anatomy (NCI_Human), (2) FMA - NCI; (3) FMA-SNOMED; (4) NCI-SNOMED. The first task corresponds to the anatomy track, and the remaining three belong to the *large biomed* track. In the *large biomed* tasks we are only aligning small overlapping fragments, which is one of the tasks supported by OAEI. This means that the portion of FMA being aligned in task 2 is not the same one that is being aligned in task 3. The same applies to NCI and SNOMED. The reference alignment used in the anatomy track was manually created and has been extensively tested. For the *large biomed* track the existing reference alignment is a silver standard based on mappings encoded in UMLS, a biomedical terminology resource [24].

5 Results and Discussion

We first evaluate an approach that uses a ranking and weighting strategy for name properties. Table 2 shows the impact on F-measure when using our proposed default ranking and weighting strategy with an interval of 0.05, in combination with two matching approaches: NM by itself, or combined with WM.

Weighting of name properties has a very noticeable impact on the alignment of the mouse and human anatomies, however that impact is much reduced in the other three matching tasks. Based on these results, and since the computational cost for this strategy is quite low, we incorporate the ranking and weighting approach into our other approaches as well.

Table 2: Ranking and weighting synonym properties.

Matchers	MA-NCI_Human	FMA-NCI	FMA-SNOMED	SNOMED-NCI
Standard AML				
NM	0.819	0.826	0.411	0.689
NM-WM	0.829	0.838	0.586	0.732
Ranking & Weighting				
NM	0.825	0.826	0.412	0.689
NM-WM	0.862	0.840	0.586	0.732

NM: name-based matcher; WM: word-based matcher

Comparison of the F-measure obtained in all four tasks when using the ranking and weighting strategy with two different matching approaches, one based on matching the full name and the other also considering word matches.

Our second approach extends the number of synonyms in ontologies either through a synonym derivation technique based on internal ontology synonyms or on the removal of common words (described in Algorithms 1, 2, and 3). These we consider to be internal synonym extension strategies, since they only use information contained in the ontologies that are being aligned. However, external ontologies can also be used to increase the number of synonyms through the transference of names from cross-referenced classes. As a source for cross-references we use the UBERON ontology, which encodes direct cross-references to the mouse and human anatomies, as well as NCI. Figure 1 shows the increase in number of name properties in each ontology after synonym extension. The number of new name properties created by intra- and inter-ontology synonym derivation is closely tied to the original number of synonyms (see Table 1), therefore for SNOMED the use of intra-ontology synonym extension does not lead to a noticeable increase in number of name properties, since there are very few synonyms to leverage on to create the internal thesaurus. However, when ontologies have very frequent words in their terminology, the number of synonyms created by the common word removal approach increases. This is clearly exemplified by SNOMED, where the existence of many names with common words such as 'structure' (e.g., 'structure of hair of trunk', 'portal vein structure' and 'spinal nerve structure') results in the creation of many more synonyms.

To test the impact of synonym extension we couple it with the NM matcher. Both intra- and inter-ontology synonym derivation can lead to a high number of erroneous names, however when used with the NM matcher these is-

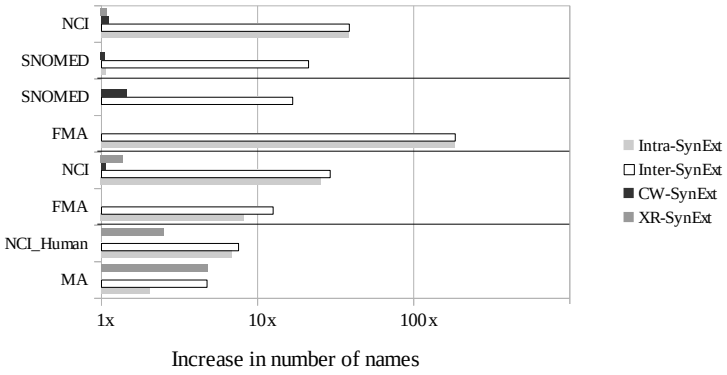

Fig. 1: Increase in number of names after synonym extension approaches for each ontology in each task.

(Intra-: intra-ontology; Inter-: inter-ontology; CW-: common word removal; XR-: cross-references to UBERON; SynExt: synonym extension)

sues are circumvented since a single match between two names is enough to map two classes and the presence of erroneous words in the names has no impact. Given the low impact intra- and inter-ontology synonym derivation has on SNOMED's terminology, we would expect a reduced impact of these strategies on the matching performance of SNOMED alignments, particularly when using the inter-ontology approach. Indeed, in the last two tasks (see Table 3), FMA-SNOMED and SNOMED-NCI have an equivalent or reduced performance when using this approach. In particular, extending SNOMED with inter-ontology synonyms leads to a marked drop in precision. On the other hand, for the alignment of the mouse and human anatomies, synonym derivation improves performance through an increase in recall, particularly for the intra-ontology approach where recall increases by 7.5%. In FMA-NCI, there is also an improvement, though not as marked, with recall increasing by 1.7%. The common word removal synonym extension approach has little to no impact on the MA-NCI_Human and FMA-NCI alignments, but has a considerable impact on FMA-SNOMED, where it increases recall by more than 40%, increasing F-measure from 41.2% to 74.5%. This is due to the fact that removing the common words in SNOMED names results in direct matches to several FMA classes. This effect is less noticeable in SNOMED-NCI, but it still increases recall by nearly 5%.

Our third approach is based on exploring external ontologies that contain cross-references to the ontologies that are to be matched, or whose domains are closely related. In this evaluation we use three ontologies as external resources: UBERON, FMA and NCI. FMA and NCI versions correspond to the full ontologies (obtained from OBO, not from OAEI). Table 4 presents the results of several distinct matching strategies that use these external ontologies. Using a combination of NM and a mediating matcher (MM) based on NM to FMA (NM-MM),

Table 3: Impact of internal synonym extension approaches on matching performance.

	No SynExt	Inter-SynExt	Intra-SynExt	CW-SynExt
	MA-NCI-human			
Precision	0.985	0.983	0.966	0.985
Recall	0.691	0.709	0.766	0.691
F-measure	0.825	0.835	0.860	0.825
	FMA-NCI			
Precision	0.945	0.936	0.939	0.944
Recall	0.723	0.736	0.74	0.723
F-measure	0.826	0.83	0.834	0.827
	FMA-SNOMED			
Precision	0.953	0.926	0.945	0.897
Recall	0.178	0.182	0.180	0.618
F-measure	0.412	0.411	0.413	0.745
	NCI-SNOMED			
Precision	0.97	0.888	0.965	0.967
Recall	0.489	0.477	0.497	0.537
F-measure	0.689	0.651	0.693	0.721

(Intra-: intra-ontology; Inter-: inter-ontology; CW-: common word removal; SynExt: synonym extension)

results in a better performance in the mouse and human anatomies as well as in SNOMED-NCI. The same strategy using NCI only impacts SNOMED-NCI results. However, when UBERON is used, there is a marked improvement in both MA-NCI_Human and FMA-NCI, which is due to MA, NCI_Human, FMA and UBERON sharing the same domain (anatomy).

UBERON encodes cross-references to MA, NCI, SNOMED and FMA. However, the cross-references are established using alphanumeric identifiers, which are unavailable in the OAEI versions of FMA and SNOMED. Consequently, we have only explored the cross-references to MA and NCI. For the MA-NCI_Human, given that UBERON encodes cross-references to both ontologies it is possible to create an alignment based solely on them (XRM). This has an F-measure of 91.7%, which is higher than any of the other approaches tested so far. A combination with NM further increases F-measure up to 92.6%. However, the cross-references can also be explored to extend the name properties of classes and then be used on an NM matching approach (NM-XR-SynExt), pushing F-measure up by another 0.9%. Combining this approach with the more complex WM results in an F-measure of 93.7% (NM-XR-SynExt-WM). The synonym extension that is based on cross-references can also be used in the NCI matching tasks, which yields the best performance we obtained for FMA-NCI, 86.4%, but has no impact on SNOMED-NCI. This is likely due to the fact that the NCI fragment in FMA-NCI belongs to the anatomy domain (the same as UBERON), whereas the SNOMED and NCI fragments of NCI-SNOMED do not.

The overall very positive success of exploring cross-references, both for direct

matching and for synonym extension, clearly demonstrates the untapped potential of these ontology properties.

Table 4: Using External Ontologies through cross-references and matching.

Matchers	MA-NCI_H.	FMA-NCI	FMA-SNM	SNM-NCI	Ext. Ont.
NM-MM	0.837	0.826	0.412	0.691	FMA
	0.826	0.827	0.412	0.691	NCI
	0.910	0.849	0.412	0.690	UBERON
XRM	0.917	N.A.	N.A.	N.A.	
+NM	0.926	N.A.	N.A.	N.A.	
NM-XR-SynExt	0.935	0.864	N.A.	0.690	
+MM	0.936	N.A.	N.A.	N.A.	
+WM	0.937	N.A.	N.A.	N.A.	

Comparison of the F-measure obtained when using different matching techniques and external ontologies to support matching. (XRM: cross-references matcher; MM: mediating matcher; WM: word-based matcher; XR-SynExt: cross-references based synonym extension)

To complete our evaluation we present a table with the comparison of our best results with the best results obtained by OAEI 2012 competitors in each task (see Table 5). For simplicity we name the integration of our approaches into AML as AMLnym. Our best results are obtained using two distinct strategies: for the MA-NCI_Human and FMA-NCI tasks the two lexical matchers (name-based and word-based) are coupled with the synonym extension derived from UBERON cross-references (NM-WM-XR-SynExt), whereas for the FMA-SNOMED and SNOMED-NCI they are coupled with the common word removal synonym extension (NM-WM-CW-SynExt). The only task where we surpass the best OAEI competitor is the MA-NCI_Human, where the use of cross-references to extend the name properties has a positive impact on performance, with 93.7% in F-measure, which is 1.4% higher than the top ranked system GOMMA-bk. For the other three tasks our results are below those obtained by the leading systems. However, both GOMMA-bk and LogMapnoe use UMLS as an external resource. Since the reference alignment is a silver standard based on UMLS, using the same resource is a biased approach that clearly results in improved performance. Considering this, we also include in the table the results obtained by those systems when using their less elaborate variants, which do not use UMLS. In FMA-NCI we still remain below GOMMA's results by 2.9%, but in the remaining tasks our approaches have a better performance, with an advantage of 24.3% in FMA-SNOMED over GOMMA and 3.7% in SNOMED-NCI over LogMapLt. However, it is important to note that GOMMA and LogMapLt differ from their more complete variants in more than just the use of UMLS, which can also explain part of the drop in performance.

Table 5: Comparison of our approaches with the best OAEI 2012 competitors in each task.

			MA-NCI_Human	FMA-NCI	FMA-SNOMED	NCI-SNOMED
OAEI 2012	AML+Nym		NM-WM-XR-SynExt		NM-WM-CW-SynExt	
		P	0.957	0.940	0.870	0.925
		R	0.917	0.802	0.670	0.589
		F	0.937	0.869	0.763	0.738
	UMLS		GOMMA-bk	GOMMA-bk	GOMMA-bk	LogMapnoe
		P	0.917	0.914	0.826	0.893
		R	0.928	0.922	0.912	0.659
		F	0.923	0.918	0.886	0.758
	no UMLS			GOMMA	GOMMA	LogMapLt
		P		0.945	0.834	0.938
		R		0.856	0.377	0.560
		F		0.898	0.520	0.701

Comparison of the performance obtained by our approaches (AML+Nym) with the best competitors in OAEI 2012 (GOMMA and LogMap) with and without the use of UMLS as an external resource (P:Precision; R:Recall; F:F-measure; NM: name-based matcher; NM: name-based matcher; WM: word-based matcher; XR-SynExt: cross-references based synonym extension; CW-SynExt: common word removal synonym extension)

6 Conclusions

We have presented three novel approaches for a better use of an ontology's terminological properties within ontology matching tasks. These approaches capitalize on biomedical ontology properties such as a rich terminology, with several synonyms of different kinds being encoded, as well as the existence of related ontologies with overlapping domains.

Our first approach distinguished between different name properties by assigning to them weights that reflect their closeness in meaning to the main name. Our results demonstrate the success of this strategy, which resulted in an increase in performance for several terminological-based matchers. Furthermore, we have shown that it is possible to extend the number of name properties of an ontology through two synonym derivation techniques, one which explores the reflexive property of synonyms to infer synonymy between words or multi-word terms that belong to synonym labels, and used these terms to compose new synonym labels, and another based on common word removal. In many cases these approaches increase the performance of name and word-based matchers up to competitive levels with more complex strategies based on external resources and structural approaches. However, the success of the synonym derivation technique based on synonym terms depends on the existence of synonyms encoded by the ontologies, which is why it is less suited for ontologies with few synonyms such as

SNOMED. The synonym derivations techniques can be also be used for ontology extension, since they are able to add novel synonyms to an ontology. Ontology extension in the biomedical domain is a budding field [25, 26], for which ontology matching has been identified as a crucial technique [27–29]. Finally, our third approach consisted in using ontologies with cross-references to the ontologies being aligned. This was shown to have a high impact on matching performance, both when the strategy was used to directly produce matches, and when it was used to extend the number of synonyms within ontologies.

The application of these approaches to OAEI tasks demonstrated the impact they can have on ontology matching performance. In the anatomy track, our results were better than those obtained by the best OAEI 2012 participant. In the three tasks of the *large biomed* track, our strategies proved insufficient to place above the leading systems. However, these systems benefit strongly from using UMLS as an external resource, and also from structural and logic-based strategies. When we compare our results with simpler versions of the leading systems that do not use these additional strategies, our approaches produce the best results in two out of three tasks. These results lead us to believe that the integration of our approaches in more complex matching strategies, using both structural and logic-based matchers will lead to an improvement of the current state of the art in biomedical ontology matching.

Furthermore, our results demonstrate that when there is an adequate external resource that links both ontologies, using it as a source for synonym extension can strongly improve matching performance. Ascertaining if an external resource is relevant for a matching task is then a relevant question, which we will address in future work. We also hope to address the extension of our synonym derivation technique to other kinds of relations such as hypernymy and holonymy.

We have demonstrated the importance of an adequate consideration of terminological properties in ontology matching, specifically of distinguishing between different synonym properties and of extending synonyms based both on ontology internal knowledge and on references to external resources. Our novel approaches will become increasingly relevant as ontologies grow and become more refined, defining more synonyms through distinct properties. We envision that the next step in exploring synonyms in biomedical ontology matching will include finding other kinds of relations, not just equivalence, so as to enable linking different entities such as diseases, symptoms, genes, anatomical structures, phenotypes and organisms, in a true biomedical Semantic Web.

Acknowledgments

DF, CP, ES and FMC were funded by the Portuguese FCT through the SOMER project (PTDC/EIA-EIA/119119/2010) and the multi-annual funding program to LASIGE. IFC and CS were partially supported by NSF Awards IIS-0812258, IIS-1143926, and IIS-1213013. IFC was also supported by a University of Illinois Scholar Award, by a UIC Area of Excellence Award, and by a UIC-IPCE Civic Engagement Research Fund Award.

References

1. Rubin, D.L., Shah, N.H., Noy, N.F.: Biomedical ontologies: a functional perspective. Briefings in bioinformatics **9**(1) (January 2008) 75–90
2. Euzenat, J., Shvaiko, P.: Ontology matching. Volume 18. Springer Berlin (2007)
3. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L., Eilbeck, K., Ireland, A., Mungall, C., et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology **25**(11) (2007) 1251–1255
4. Noy, N., Shah, N., Whetzel, P., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D., Storey, M., Chute, C., et al.: Bioportal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research **37**(suppl 2) (2009) W170–W173
5. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., Trojahn, C.: Ontology alignment evaluation initiative: six years of experience. Journal on data semantics XV (2011) 158–192
6. Ghazvinian, A., Noy, N., Musen, M.: Creating mappings for ontologies in biomedicine: Simple methods work. In: AMIA Annual Symposium (AMIA 2009). Number 2 (2009)
7. Lambrix, P., Tan, H.: SAMBO - system for aligning and merging biomedical ontologies. Web Semantics: Science, Services and Agents on the World Wide Web **4**(3) (2006) 196–206
8. Groß, A., Hartung, M., Kirsten, T., Rahm, E.: GOMMA results for OAEI 2012. In: Ontology Matching Workshop. International Semantic Web Conference 2012 (2012)
9. Cruz, I.F., Stroe, C., Caimi, F., Fabiani, A., Pesquita, C., Couto, F.M., Palmonari, M.: Using AgreementMaker to Align Ontologies for OAEI 2011. In: ISWC International Workshop on Ontology Matching (OM). Volume 814 of CEUR Workshop Proceedings. (2011) 114–121
10. Cruz, I.F., Palandri Antonelli, F., Stroe, C.: AgreementMaker: Efficient Matching for Large Real-World Schemas and Ontologies. PVLDB **2**(2) (2009) 1586–1589
11. Xu, P., Tao, H., Zang, T.: Alignment Results of SOBOM for OAEI 2009. In: The 4th International Workshop on Ontology Matching at ISWC 2009
12. Jiménez-Ruiz, E., Grau, B.C., Zhou, Y.: LogMap 2.0: towards logic-based, scalable and interactive ontology matching. In: Proc. of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences. (2011) 45–46
13. Aguirre, J.L., Eckert, K., Euzenat, J., Ferrara, A., van Hage, W.R., Hollink, L., Meilicke, C., Nikolov, A., Ritze, D., Scharffe, F., Shvaiko, P., Sváb-Zamazal, O., dos Santos, C.T., Jiménez-Ruiz, E., Grau, B.C., Zapilko, B.: Results of the Ontology Alignment Evaluation Initiative 2012. In: ISWC International Workshop on Ontology Matching (OM). Volume 946 of CEUR Workshop Proceedings. (2012) 73–115
14. Cruz, I.F., Sunna, W., Makar, N., Bathala, S.: A Visual Tool for Ontology Alignment to Enable Geospatial Interoperability. Journal of Visual Languages and Computing **18**(3) (2007) 230–254
15. Morant, A.: Extending and optimizing an ontology matching system (2011)
16. Couto, F., Silva, M., Coutinho, P.: Finding genomic ontology terms in text using evidence content. BMC bioinformatics **6**(Suppl 1) (2005) S21
17. Gross, A., Hartung, M., Kirsten, T., Rahm, E.: Mapping composition for matching large life science ontologies. In: ICBO. (2011)

18. Faria, D., Pesquita, C., Santos, E., Cruz, I.F., Couto, F.M.: The AgreementMakerLight Ontology Matching System. In: ODBASE. (2013)
19. Cruz, I.F., Sunna, W.: Structural Alignment Methods with Applications to Geospatial Ontologies. Transactions in GIS, Special Issue on Semantic Similarity Measurement and Geospatial Applications **12**(6) (2008) 683–711
20. Euzenat, J., Ferrara, A., Hollink, L., Isaac, A., Joslyn, C., Malaisé, V., Meilicke, C., Nikolov, A., Pane, J., Sabou, M., Scharffe, F., Shvaiko, P., Spiliopoulos, V., Stuckenschmidt, H., Sváb-Zamazal, O., Svátek, V., dos Santos, C.T., Vouros, G.A., Wang, S.: Results of the Ontology Alignment Evaluation Initiative 2009. In: ISWC International Workshop on Ontology Matching (OM). Volume 551 of CEUR Workshop Proceedings. (2009) 73–126
21. Euzenat, J., Ferrara, A., Meilicke, C., Pane, J., Scharffe, F., Shvaiko, P., Stuckenschmidt, H., Sváb-Zamazal, O., Svátek, V., dos Santos, C.T.: Results of the Ontology Alignment Evaluation Initiative 2010. In: ISWC International Workshop on Ontology Matching (OM). Volume 689 of CEUR Workshop Proceedings. (2010) 85–117
22. Euzenat, J., Ferrara, A., van Hage, W.R., Hollink, L., Meilicke, C., Nikolov, A., Ritze, D., Scharffe, F., Shvaiko, P., Stuckenschmidt, H., Sváb-Zamazal, O., dos Santos, C.T.: Results of the Ontology Alignment Evaluation Initiative 2011. In: ISWC International Workshop on Ontology Matching (OM). Volume 814 of CEUR Workshop Proceedings. (2011) 85–113
23. Ngo, D., Bellahsene, Z.: Yam++: A multi-strategy based approach for ontology matching task. In: Knowledge Engineering and Knowledge Management. Springer (2012) 421–425
24. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res **32**(Database issue) (January 2004)
25. Pesquita, C., Couto, F.: Predicting the extension of biomedical ontologies. PLOS Computational Biology **8**(9) (2012) e1002630
26. Wächter, T., Schroeder, M.: Semi-automated ontology generation within OBO-Edit. Bioinformatics **26** (2010) 88–96
27. Novácek, V., Laera, L., Handschuh, S., Davis, B.: Infrastructure for dynamic knowledge integration–automated biomedical ontology extension using textual resources. Journal of biomedical informatics **41**(5) (October 2008) 816–28
28. Pesquita, C., Stroe, C., Cruz, I.F., Couto, F.: BLOOMS on AgreementMaker: Results for OAEI 2010. In: ISWC International Workshop on Ontology Matching (OM). Volume 689 of CEUR Workshop Proceedings. (2010) 135–141
29. Pesquita, C.: Automated extension of biomedical ontologies. PhD thesis, University of Lisbon (2012)