Complex multi-ontology alignment through geometric operations on language embeddings

Marta Contreiras Silva[a,*], **Daniel Faria**[b] **and Catia Pesquita**[a]

[a]LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
[b]INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal
ORCID (Marta Contreiras Silva): https://orcid.org/0000-0003-1864-0105, ORCID (Daniel Faria): https://orcid.org/0000-0003-1511-277X, ORCID (Catia Pesquita): https://orcid.org/0000-0002-1847-9393

Abstract.

With knowledge graphs increasing in popularity, aligning and integrating them is paramount to ensure their usefulness and reusability. A key step in this process is ontology matching, whereby the semantic models of KGs are aligned into a single cohesive semantic backbone. While finding simple pairwise equivalences between entities in two ontologies is well addressed by state-of-the-art algorithms, finding more complex mappings that can include multiple entities from different ontologies is far from solved, despite their importance in ensuring a deep and meaningful integration of KGs.

We propose a novel complex ontology matching approach that explores geometric operations over the shared semantic space afforded by large language models, enabling the discovery of complex mappings that are missed by purely lexical approaches. We evaluate our approach on several biomedical ontologies using partial reference alignments and manual expert validation. Our approach improves on the performance of a purely lexical approach while also increasing the coverage of complex multi-ontology alignments by 20 to 80%, which translates to a 97% coverage of the source ontologies. Moreover, the manual evaluation of the mappings produced by LLM shows that it achieves a high level of precision. This work demonstrates that the use of LLMs can improve on the performance of traditional lexical strategies.

1 Introduction

Knowledge Graphs (KGs) imbue data with semantics which both facilitates and augments information retrieval and knowledge discovery [23]. As the semantic model of a KG is often composed of one or more ontologies, ontology matching—the process of finding mappings between related ontologies—is essential both for its construction and enabling semantic interoperability between data modelled under different KGs.

Traditionally, ontology matching algorithms have focused almost exclusively on finding simple equivalence mappings between single entities of two ontologies [6, 16, 7]. However, these mappings will rarely suffice for achieving complete semantic integration between two ontologies, even within the same exact domain, as they will differ to some extent on how they view and model that domain [25]. For example, one ontology might have a richer class hierarchy whereas another relies more heavily on properties. Thus, the first ontology declares class *S:predator* as a subclass of *S:animal*, and in the second ontology being a predator can only be asserted through *T:animal* plus hunting an animal (where *T:hunts* is an object property and *T:animal* is a class). Through a simple mapping, we can only capture that *S:predator* is narrower than class *T:animal*, but using a complex mapping we can detail the precise relationship between these aspects to enable full semantic integration: *S:predator* $\equiv$ *T:animal* AND (*T:hunts* SOME *T:animal*). Yet finding complex mappings is highly challenging, as evidenced by the fact that the performance of state-of-the-art tools is underwhelming in comparison with their performance in simple matching tasks [1].

Constructing KGs often requires matching numerous ontologies in order to achieve a holistic view of a domain [14, 27]. When restricted to finding simple mappings, the multi-ontology matching problem has typically been broken down into a collection of pairwise matching problems, enabling the use of established ontology matching algorithms [11, 14, 26, 27]. But to achieve full integration of ontologies, we must also contemplate complex mappings—including mappings that contain expressions combining entities of two or more ontologies—especially if we want to combine ontologies of deeply related but disjoint domains (e.g., human phenotypes, phenotypic qualities, and human anatomy) to get a holistic view of their super-domain (e.g., biomedicine). Complex multi-ontology mappings are commonly found in biomedical ontologies, under the name of logical definitions, as a result of a concerted effort from ontologists to achieve this type of semantic integration [20]. These logical definitions are always 1:n complex mappings, as they connect classes of the ontology in which they are present to expressions combining entities from one or more target ontologies, e.g.: *abnormality of glycosaminoglycan metabolism* $\equiv$ (*quality* AND (*characteristic_of_part_of* SOME *glycosaminoglycan metabolic process*) AND (*has_modifier* SOME *abnormal*)). To date, the only notable effort at finding these types of mappings automatically achieved moderate success but tackled only the simplest instances of the problem (mappings involving only two classes from exactly two target ontologies) [22].

We propose a novel approach for tackling Complex Multi-Ontology Matching (CMOM) in a 1:n setting without restrictions on the arity of the mappings or on the number of ontologies, which we call Complex Multi-Ontology Matching through Recursive Sub-

[*] Corresponding Author. Email: mcdsilva@fc.ul.pt

traction (CMOM-RS)[1]. CMOM-RS combines two recursive strategies for candidate mapping generation: a strategy based on Large Language Models (LLM) that exploits the semantic space conveyed by embeddings through geometric operations; and a lexical strategy that relies on string similarity. These two strategies complement each other, as the lexical strategy is precise but restricted to cases of perfect syntactic matches, whereas the LLM-based strategy is less precise but provides much greater coverage. We evaluate these strategies extensively using partial reference alignments extracted from biomedical ontologies, as well as through a manual expert evaluation of a sample of mappings for classes not in the reference.

The contributions of this work are CMOM-RS, a first approach to CMOM with no arity restrictions, a first implementation of semantic precision and recall metrics to the CMOM setting, and a set of partial reference alignments for evaluating CMOM.

2 Problem Definition

In its original form, Ontology Matching aims to find mappings (or correspondences) relating an entity of one source ontology with an entity of one target ontology. A mapping is usually represented as a tuple of the form $< e_1, e_2, r, c >$ [6], where e_1 and e_2 are the entities of the source and target ontology, respectively, r is the semantic relation between them (e.g. $\equiv$, $\geq$, $\leq$, $\perp$), and c is an optional confidence score. The mapped entities can be classes, individuals, or properties, and a set of mappings between two ontologies is called an alignment.

Complex Ontology Matching is an extension of the original ontology matching paradigm whereby (complex) mappings can include logical expressions—comprising one or more entities of the two ontologies—in place of one or both e_1 and e_2 [25], to enable the capturing of more precise semantic relations between the two ontologies. These logical expressions commonly include intersection or union of classes, as well as restrictions on the occurrence, domain, range or cardinality of properties. Finding complex mappings is much more challenging than finding simple mappings, as it involves linking several entities in a mapping and also determining the logical constructs with which to combine them. Acknowledging this challenge, some evaluation efforts focus simply on whether algorithms can identify the mapped entities, rather than the exact expressions in the mappings, under the rationale that a human curator can easily compose the expressions if given the correct entities, but finding the entities manually would be very time-consuming [30].

Multi-Ontology Matching is another extension to the original ontology matching paradigm, where the goal is to integrate N>2 ontologies through simple mappings, often contemplating only equivalence relations [19]. As the problem is typically broken down into a collection of pairwise matching problems [11, 14, 26, 27], mappings are still represented in the same way as in simple ontology matching, although they may be merged into a final alignment wherein e_1 and e_2 can be entities of any of the N matched ontologies.

CMOM, as the name suggests, contemplates complex mappings in a multi-ontology setting, including complex mappings with expressions comprising entities of several ontologies. In its broadest, m:n form, we can formally define CMOM as the task of finding mappings of the form $< e_1, e_2, r, c >$, where e_1 and e_2 can be entities or expressions comprising any number of distinct entities from any number of ontologies. However, the applicability of m:n multi-ontology mappings to real world scenarios is narrow, whereas 1:n multi-ontology mappings are commonly present in biomedical ontologies. Thus, this work focuses on 1:n CMOM, where e_1 is an entity of a single source ontology, and e_2 is an expression of any number of distinct entities from any number of target ontologies.

Even in its 1:n form, the CMOM problem is nearly untractable, as the search space is boundless: there are no theoretical restrictions on the arity of the target expression, which can include any number of distinct entities of any number of ontologies, combined in any manner conceivable within the expressivity limits of the OWL language (which allows boundless nesting of expressions). Even if we restrict the search space by placing plausible limits on the maximum arity of the target expression (11 in existing logical definitions) and on the range and maximum nesting of logical constructs it can include (3 types of constructs with typically 2-3 layers of nesting in existing logical definitions) the number of combinations is still astronomical. To reduce the complexity of the problem, we focus our effort on identifying only the mapped entities, as has been proposed for pairwise complex matching [30].

3 Related Work

The field of ontology matching is dominated by algorithms and tools for tackling the simple matching problems. The most successful among these tools rely on rule-based algorithms primarily based on the lexical component of ontologies (i.e., string matching algorithms) while also contemplating the structural component and exploiting sources of background knowledge [7, 16].

Efforts at tackling complex matching in a pairwise setting are relatively few, and can be divided into [28]: lexical-based strategies, which rely on finding partial lexical overlaps between entities; and instance-based strategies, which rely on pattern analysis of the classes and properties associated with instances shared between two ontologies. The latter strategies are more robust than the former, as they depend on the usage of the ontologies rather than their terminology, and thus are not affected by terminological differences between ontologies. However, they are limited to tackling problems for which shared instances exist, which largely precludes their applicability to the CMOM scenario, where the ontologies to integrate are typically from related but disjoint domains.

The only notable effort at tackling CMOM is, to the best of our knowledge, the work of Oliveira *et al.* [22], who adapted AgreementMakerLight [7] to finds ternary mappings (between one source and two target ontologies) of specific patterns, through a lexical-based strategy enhanced by a word stemmer. It should be noted that their approach was restricted to the simplest case within the CMOM scenario. To date, no strategy has been proposed to tackle CMOM with no restrictions on the arity of the mappings.

The recent emergence of LLMs has not gone unnoticed in ontology matching, as their applicability to a field that is largely based on lexical strategies is readily apparent. As of the latest edition of the Bio-ML track[2] of the Ontology Alignment Evaluation Initiative[3], five systems reported the use of LLMs to tackle simple ontology matching tasks: AgreementMakerDeep [29], BERTMap [13], Matcha-DL [8], OLala [15], and SORBETMatcher [10]. All five systems rely on the cosine similarity between the LLM embedding representations of the labels/synonyms to predict mappings, and except for OLala, all combine their LLM-based strategy with string matching. OLala and Matcha-DL also differ from the other three systems in that they do not perform fine-tuning based on the ontology corpora.

[1] https://github.com/liseda-lab/CMOM-RS

[2] https://www.cs.ox.ac.uk/isg/projects/ConCur/oaei/2023/index.html

[3] https://oaei.ontologymatching.org/2023/

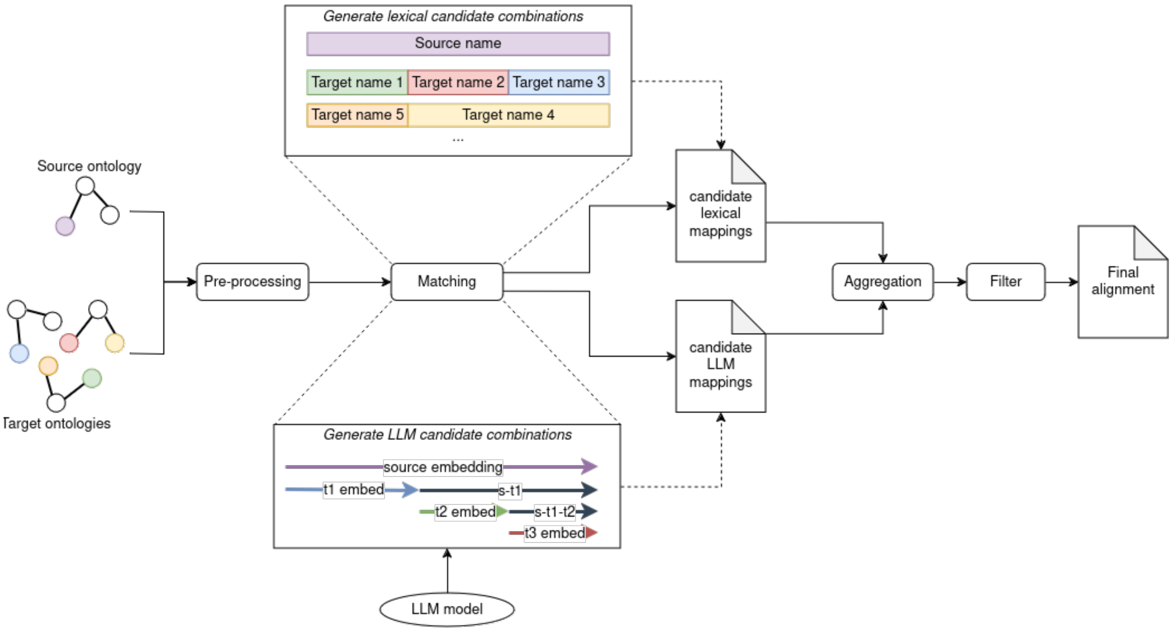

Figure 1. Overview of CMOM-RS, our approach for Complex Multi-ontology Ontology Matching. One source ontology and a set of target ontologies serve as input to two strategies, lexical and LLM, that produce candidate mappings. These are then aggregated and filtered to obtain the final alignment.

To date, no LLM-based strategy has been proposed for CMOM or even for the simpler pairwise complex ontology matching.

4 Methods

4.1 Overview

Our proposed approach, CMOM-RS, aims to find complex multi-ontology mappings of the form $< e_s, \{e_{t_1}, ..., e_{t_m}\}, \equiv, c >$, where S is a source ontology with $e_s \in S$ being the source entities, $e_t \in \{T_1, ..., T_n\}$ are the target entities present in a set of n target ontologies with $n \geq 1$, mappings are of cardinality $1 : m$ with $m \geq 2$, and c the confidence score. It is composed of several steps: (1) pre-processing, whereby the ontologies vocabularies are extracted and processed for further use; (2) matching, based on two complementary candidate mapping generation approaches (one lexical and one that explores LLM representations of entities); (3) aggregation and filtering, where candidate mappings are selected to compose a final alignment. A general overview of the approach is shown in Figure 1. The reasoning behind exploring two candidate mapping generation approaches is to explore in tandem the complementarity between a precision-oriented approach (lexical) and a recall-oriented approach (LLM).

4.2 Pre-processing

Both the lexical and LLM-based matching algorithms rely on the vocabulary component of the ontologies. The first step in our approach is to extract the set of labels from the S into a source vocabulary N_S, and the set of labels from $\{T_1, ..., T_n\}$ into a unified target vocabulary N_T. Each label, hereafter name, is associated to its original class, as well as with a confidence score. Leveraging on the fact that many ontologies define multiple labels for the same entity, and that some even define different types of synonyms[4], the pre-processing approach assigns a confidence score that reflects the semantics of the label property (higher for local names and lower for synonyms) and also corrects for the frequency of the label properties (see more details in Appendix section 8.1[5]).

4.3 Lexical Candidate Generation

The generation of lexical candidates uses the ontologies' vocabularies as the basis for its strategy. For each name in N_S, a filtered target vocabulary is generated by filtering N_T to remove any names that do not share at least one word with the source name. Then a recursive function is applied to find all possible combinations of target names that do not overlap with each other and afford full coverage of the source class name (see Algorithm 1 in the Appendix). Each of these combinations corresponds to a candidate mapping. Since candidates are identified through names, each source entity may have multiple names, and there may be multiple valid target combinations, meaning multiple candidates (at the name level) correspond to the same mapping (at the entity level). The aggregation of these candidates is conducted in step (3).

The confidence score of each candidate mapping M is calculated as the product of the confidence of each target name n_{t_j} present in it (see Equation 1).

$$score = \prod_{j=1}^{n} conf(n_{t_j}) \tag{1}$$

4.4 Large Language Model-based candidate generation

The LLM-based candidate generation is the core of our proposed approach. It also relies on the ontologies' vocabularies but explores

[4] *oboInOwl:hasRelatedSynonym*, *oboInOwl:hasExactSynonym*, *oboInOwl:hasBroadSynonym*, and *oboInOwl:hasNarrowSynonym*

[5] Appendix: https://github.com/liseda-lab/CMOM-RS

latent representations generated by encoder language models to capture the distributional semantics of the entity names through geometric operations: vector subtraction and cosine similarity (see Algorithm 2 in the Appendix). Unlike the lexical candidate generation, which is able to generate multiple candidates for the same source entity name, the LLM-based candidate generation aims to find the best set of target names to compose a mapping for a source class name s (FIND BEST MAPPING(s) in Algorithm 2). It relies on a recursive approach that finds the most similar target name embedding to the source embedding, updates the source embedding by subtracting the target embedding from it, and recursively finds the next most similar target embedding until the cosine similarity between the two embeddings is lower than the input parameter α (Figure 2). At each iteration, the most similar target name found is added to the mapping.

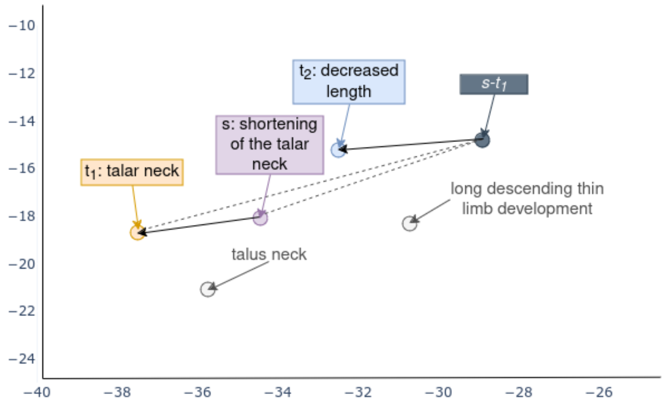

Figure 2. Visualisation of the first two steps of the construction of a complex mapping in a 2D space using a geometric operation.

More formally, let s be the source name embedding, then the best mapping for s can be found following Equations 2.

$$map(s) = \begin{cases} M = M \cup t_{max} & \text{if } cos(s, t_{max}) < \alpha \\ map(s - t_{max}) & \text{if } cos(s, t_{max}) \geq \alpha \end{cases} \tag{2}$$

where M is the set of target names selected to compose the best mapping, $s - t_{max}$ corresponds to subtracting the vectors s and t_{max} is the embedding for the most similar target name, computed by maximizing the cosine similarity between the source name embedding and the target names embeddings t_i (see Equation 3).

$$t_{max} = \underset{\mathbf{t}_i}{\text{argmax}}\, cos(s, t_i) \tag{3}$$

The inspiration for the use of subtraction is the intuition that a composition of target vectors would approximate the source vector[6].

The confidence score for the LLM-based candidates is calculated as the cosine similarity between the source name embedding s and the sum of the target name embeddings t_j used in the candidate mapping (see Equation 4).

$$\text{score} = \cos(\mathbf{s}, \sum_{j=1}^{n} \mathbf{t}_j) \tag{4}$$

4.5 *Aggregation and Filtering*

In order to aggregate both sets of candidates into a single alignment and ensure a cardinality of approximately 1:1, we employ a greedy heuristic to select mappings sorted by descending confidence score as long as they do not conflict with already selected mappings, to produce a (near) 1-to-1 alignment, since tied mappings are all returned instead of choosing arbitrarily between them. This approach follows the paradigm proposed in [7].

Table 1. Frequency of usage of unique entities in each target ontology in the logical definitions of the three source ontologies and number of total mappings extracted

	Source ontologies		
Target ontologies	HP	MP	WBP
ChEBI [12]	219	172	110
CL [3]	82	363	5
GO [2]	237	834	305
PATO [9]	308	369	103
UBERON [21]	1154	2093	1
WBbt [18]	-	-	136
Total mappings	5447	9311	869

5 Experimental Design

5.1 *Implementation*

CMOM-RS was implemented using a pre-trained Sentence-BERT model [24][7] that used a self-supervised contrastive learning objective that makes it especially appropriate for tasks such as sentence similarity, and therefore well-suited to ontology alignment tasks focused on vocabulary similarity. Other similar pre-trained models can be employed.

Although our reference logical definitions use both classes and properties, the properties are highly repetitive and they offer little added value to the construct—looking at examples such as *part of* or *characteristic of*—since they are disconnected from the definition of the source entity. Moreover, a human curator could effortlessly compose the expression from a set of target entities. Due to this, our approach is focused on finding the target entities involved in the complex mappings and not on finalizing a complete construct.

5.2 *Ontologies and Reference Alignments*

We evaluated our approach on three CMOM tasks extracted from the biomedical domain. These tasks were selected since they are, to the best of our knowledge, the only tasks for which a partial reference alignment can be extracted to evaluate CMOM-RS. Each of these tasks is focused on a source ontology: the Human Phenotype Ontology (HP), the Mammalian Phenotype Ontology (MP), and the Worm Phenotype Ontology (WBP). Each of these ontologies encodes logical definitions that cover a diverse set of ontologies. To build the reference alignments (see Table 1) for both the HP and MP task, we employed the following target ontologies: the Cell Ontology (CL), the Chemical Entities of Biological Interest (ChEBI), the Gene Ontology (GO), the Phenotype and Trait Ontology (PATO), and the Uber Anatomy Ontology (UBERON). For the WBP task, the target ontologies were: ChEBI, GO, PATO, and the *C.elegans* Gross Anatomy Ontology (WBbt). For a more complete analysis of the logical definitions see section 8.2 in the Appendix.

[6] This approach is reminiscent of the vector operations popularized by word2vec that were later shown to only work on a few selected cases, but which generalizes well in your task.

[7] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

5.3 Experiments

Our experimental design includes several experiments. We generated partial alignments to evaluate against the partial reference alignments, where mappings are only created for the source entities present in the reference. We also generated full alignments where all source ontology entities can be matched. A set of domain experts manually evaluated a sub-sampling of a full alignment. In addition, we also conducted an ablation study to assess the influence of the different candidate generation components.

5.4 Reference-based Evaluation Metrics

Considering the novelty of the CMOM task, we developed a novel approach to evaluate complex multi-ontology mappings that extends existing paradigms in both the semantic evaluation of ontology alignments [5] and the evaluation of complex alignments [31].

A candidate mapping for a given source entity is considered fully correct if the target entities in the reference mapping for that source entity correspond exactly to the target entities in the candidate mapping. Our evaluation approach accounts for two types of partial correctness: (1) semantic correctness, where semantically similar entities are still considered a positive contribution to the correctness of the mapping; and (2) completeness, where the presence or absence of a target entity in the candidate mapping versus the reference mapping penalizes precision or recall respectively.

Regarding semantic correctness, the relaxed precision and relaxed recall of each mapping are calculated according to Equations 5 and 6 respectively.

$$relaxed\,prec(e_i) = \begin{cases} 1 & \text{if } e_i \leq e_{rj} \\ 0.5 & \text{if } e_i > e_{rj} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$relaxed\,rec(e_i) = \begin{cases} 1 & \text{if } e_i \geq e_{rj} \\ 0.5 & \text{if } e_i < e_{rj} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where e_i and e_{rj}, stand for a predicted entity and a reference entity, respectively, and $>$ and $<$ stand for direct sub- or super-classes.

Regarding completeness, the precision of a mapping is then computed as the ratio between the sum of the relaxed precision values of each target entity in the predicted mapping and the number of target entities in the candidate mapping (Equation 7).

$$prec(m_i) = \frac{\sum_{i=1}^{n} \text{relaxed precision}(e_i)}{n} \tag{7}$$

where n stands for the number of target entities in the candidate mapping, and m_i stands for a candidate mapping composed of entities $\{e_1, .., e_n\}$.

Mapping recall is calculated using the ratio between the sum of the relaxed recall values of each individual target entity and the number of target entities in the reference mapping (Equation 8).

$$rec(m_i) = \frac{\sum_{i=1}^{n} \text{relaxed recall}(e_i)}{r} \tag{8}$$

where r stands for the number of target entities in the reference mapping.

A subsequent aggregation to evaluate the full alignment was performed for both metrics. Alignment precision is calculated as the ratio between the sum of all mapping precision values and the total number of mappings in the alignment (Equation 9).

$$\text{alignment precision} = \frac{\sum_{i=1}^{n} prec(m_i)}{|M|} \tag{9}$$

where $prec(m_i)$ stands for the value of precision of a mapping and M for the total number of mappings in the final alignment.

Similarly, the alignment recall is calculated as the ratio between the sum of all mapping recall values and the total number of mappings in the reference alignment (Equation 10).

$$\text{alignment recall} = \frac{\sum_{i=1}^{n} rec(m_i)}{|M_r|} \tag{10}$$

where $rec(m_i)$ stands for the value of recall of a mapping and M_r stands for the total number of mappings in the reference alignment.

In addition to semantic precision, recall and F1-score, we also classified each mapping according to its level of correctness and completeness into mutually exclusive and ordered categories:(i) *correct* (predicted mapping and reference mapping match in all entities); (ii) *contains* (the predicted mapping contains the reference mapping); (iii) *contained* (the predicted mapping is contained within the reference mapping); (iv) *overlap* (some entities are shared between the two mappings). Each of these four categories was extended to account for semantic correctness (marked with *). Considering the

Table 2. Example provided to evaluators for expert-based evaluation of the LLM mappings.

Source name	Source synonyms	Target name	Target synonyms	Evaluation per target	Complete?	Confidence (1-4)
abnormal basophil morphology (HP_0001912)	abnormality of basophils	basophil (CL_0000767)	basophilic leucocyte, basophilic leukocyte, polymorphonuclear leucocyte, polymorphonuclear leukocyte	correct	no	4
		morphology (CL_0000051)	-	correct		
abnormal basophil morphology (HP_0001912)	abnormality of basophils	abnormal (PATO_0000460)	atypical; atypia; aberrant; defective	correct	yes	4
		basophil (CL_0000767)	basophilic leucocyte, basophilic leukocyte, polymorphonuclear leucocyte, polymorphonuclear leukocyte	correct		
		structure (PATO_0000141)	conformation, relational structural quality	related		
abnormal basophil morphology (HP_0001912)	abnormality of basophils	abnormal (PATO_0000460)	atypical; atypia; aberrant; defective	correct	yes	4
		basophil (CL_0000767)	basophilic leucocyte, basophilic leukocyte, polymorphonuclear leucocyte, polymorphonuclear leukocyte	correct		
		morphology (PATO_0000051)	-	correct		

inherent complexity of the CMOM task and that OM is typically followed by user validation to ensure the correctness of the alignment [4], these categories allow us to delve into the effort required to correct the candidate mappings.

5.5 *Expert-based Evaluation Approach*

Since the reference alignments are only partial—they only provide mappings for a subset of source classes—we recruited three domain experts[8] (7-20 years of experience in the biomedical field) to assess the validity of the top 50 HP mappings by confidence score generated by the LLM-based approach. The mappings were presented to the experts in a table format (example in Table 2), and they were asked to evaluate both target entity correctness and full mapping completeness.

Regarding target entity correctness, each entity could be evaluated according to:

$$correctness(e_i, e_r) = \begin{cases} 1 & \text{if } e_i \supset e_r \quad (correct) \\ 0.5 & \text{if } e_i \sim e_r \quad (related) \\ 0 & \text{if } e_i \not\supset e_r \quad (incorrect) \end{cases} \tag{11}$$

Mappings whose target entities are correct or related can also be evaluated as complete or incomplete, in order to assess whether the set of target entities achieves full coverage of the source entity's meaning. We approximate incompleteness to $n + 1$, where n is the number of target entities in the candidate mapping:

$$completeness(E_t, e_r) = \begin{cases} n & \text{if } E_t, \equiv e_r \quad (complete) \\ n+1 & \text{if } E_t \supset e_r \quad (incomplete) \end{cases} \tag{12}$$

where E_t is the set of target entities in a candidate mapping m (i.e., $\{e_{t_1}$,..., $e_{t_n}\}$).

These scores can be used to calculate manual precision and recall values for each mapping. Precision is calculated as the ratio between the sum of the correctness scores for each target entity and the number of target entities in the candidate mapping n. For recall, we considered the ratio between the sum of the correctness scores for each target entity and the completeness score of the candidate mapping (Equations 13 and 14 in the Appendix). The alignment evaluation is then computed using Equations 9 and 10.

Expert evaluators were free to search for any additional information they deemed necessary (e.g., ontology browsers) to conduct the evaluation and were asked to grade their confidence in each mapping evaluation on a scale from 1 to 4.

The Cohen's κ coefficient was calculated to assess the agreement between experts for both metrics measured. We employed a quadratically weighted κ to account for the ordinal aspect of the scale for correctness, i.e., 'correct' and 'related' evaluations are considered to agree more than 'correct' and 'incorrect' ones.

Table 3. Results for the reference-based evaluation.

		Semantic Precision	Semantic Recall	Semantic F1-score
HP	CMOM-RS w/o LLM	**0.620**	0.392	0.481
	CMOM-RS w/o lexical	0.472	0.466	0.469
	CMOM-RS	0.510	**0.579**	**0.543**
MP	CMOM-RS w/o LLM	**0.760**	0.643	**0.696**
	CMOM-RS w/o lexical	0.559	0.511	0.534
	CMOM-RS	0.651	**0.722**	0.685
WBP	CMOM-RS w/o LLM	**0.617**	0.391	0.479
	CMOM-RS w/o lexical	0.407	0.456	0.430
	CMOM-RS	0.461	**0.520**	**0.489**

6 Results and Discussion

Tables 3 and 4 present the results of our reference-based evaluation. Due to the lack of established baselines in CMOM, we compared our CMOM-RS approach with two ablated variants: without the lexical candidate generation (only LLM) and without the LLM-based candidate generation (only lexical, which can be seen as an improved version of the state-of-the-art method established by [22]). Table 3 shows that the complete CMOM-RS approach achieves the best semantic recall for all tasks and the best F1-score for HP and WBP. The ablated version without the LLM-based candidate generation achieves the best F1-score for MP due to its higher precision.

The higher precision of the lexical candidate generation in comparison with the LLM-based candidate generation can be explained by the fact that in the former we enforce the source entity to be the disjoint union of the target entities syntactically, whereas in the latter we can only approximate this criterion in the semantic space conveyed by the LLM embeddings. As the ontologies involved are of closely related domains, the syntactical identity of the lexical candidate generation translates well to semantic identity, whereas the semantic latitude afforded by the LLM embeddings leads to less precise mappings. On the other hand, the LLM-based candidate generation has a higher recall than the lexical candidate generation in two of the three tasks, and contributes substantially to the recall of the complete strategy in all three, as it is able to find (partially) correct mappings in cases where there is semantic but not syntactical alignment between the entities.

[8] who are not authors of this paper

Table 4. Results for the reference-based evaluation of mapping correctness and completeness, according to the categories of correct, contains, contained and overlap. Categories marked with * aggregate mappings that are constructed using a direct sub- or super-class.

		Mappings	Source entities	Target entities	Correct	Correct*	Contains	Contains*	Contained	Contained*	Overlap	Overlap*	Total
HP	CMOM-RS w/o LLM	2967	2596	2947	710	63	140	248	71	31	1273	195	2731
	CMOM-RS w/o lexical	5857	5381	5825	609	162	46	40	513	62	2195	445	4072
	CMOM-RS	5987	5406	5967	826	105	156	257	393	59	2366	438	4600
MP	CMOM-RS w/o LLM	7521	6867	7481	2328	151	678	1025	419	49	2199	423	7272
	CMOM-RS w/o lexical	9908	9223	9860	1677	170	25	3	1329	152	3868	407	7631
	CMOM-RS	10122	9282	10080	2441	171	689	1027	543	80	3131	598	8680
WBP	CMOM-RS w/o LLM	542	480	542	53	45	100	16	13	4	206	26	463
	CMOM-RS w/o lexical	993	862	993	48	36	4	2	20	6	522	57	695
	CMOM-RS	962	863	962	56	47	101	17	14	4	401	48	688

Table 5. Statistics for both the reference and non-reference alignments, as well coverage against the total classes in the ontology.

			Reference		Full		
		Classes	Mappings	Source entities	Mappings	Source entities	Coverage
HP	Original LDs	16601	5477	5447	-	-	36.99%
	CMOM-RS w/o LLM		2967	2596	3905	3468	25.06%
	CMOM-RS w/o lexical		5857	5381	17087	15506	97.58%
	CMOM-RS		5987	5406	17197	15538	97.77%
MP	Original LDs	14513	9311	9311	-	-	69.83%
	CMOM-RS w/o LLM		7521	6867	8632	7792	59.37%
	CMOM-RS w/o lexical		9908	9223	14469	13482	98.57%
	CMOM-RS		10122	9282	14806	13565	99.15%
MP	Original LDs	2701	869	869	-	-	35.69%
	CMOM-RS w/o LLM		542	480	977	862	35.43%
	CMOM-RS w/o lexical		993	862	2893	2562	98.37%
	CMOM-RS		962	863	2866	2571	98.70%

Table 4 shows that our approach is able to generate more mappings than either ablated version, but also more fully correct mappings, demonstrating that aggregating both candidate generation methods is beneficial. It is also interesting to note that 8% to 28% of the mappings found by CMOM-RS are fully correct, and between 61 and 65% are partially correct, amounting to a relatively low effort in a manual validation scenario.

As any evaluation using a partial reference alignment is, by definition, incomplete, Table 5 shows the global statistics of not only the partial alignments (i.e., only for source classes with logical definitions) but also the full alignments generated by CMOM-RS and its two ablation variants. The results show that CMOM-RS is close to full coverage of all three source ontologies, indicating it can be used to find logical definitions for several thousand classes that currently lack them.

While in principle, the performance of CMOM-RS could be extrapolated from the partial to the full alignments, we also performed a manual evaluation of a sample of 50 mappings for source classes with no logical definitions. The evaluators classified each target as correct, related or incorrect—equal to 1, 0.5, and 0 respectively—and each mapping as complete or incomplete. The aggregated results can be seen in Table 6.

Table 6. Manual evaluation results.

Correctness Cohen's *k*	Completeness Cohen's *k*	Precision	Recall
0.705	0.592	0.880	0.851

The agreement of the evaluators responses is presented as the Cohen's kappa coefficient, calculated as the average of the coefficient between all evaluator pairs. The responses indicate that evaluators agreed more often on the evaluation of the targets than whether or not the mapping was complete, with target classification results being 'substantial' and mapping completeness being 'moderate' [17]. 77% of the mappings were evaluated with the maximum confidence score by the experts.

The accuracy of the manually evaluated mappings is high, with an average precision of 88% and an average recall of 85%, meaning the mappings were fairly close to correct according to the experts. Although these scores are higher than the semantic precision and recall scores observed for the reference-based evaluation, the two sets of scores are not directly comparable.

Mappings that were classified as complete and correct by all evaluators will be made available as potential new reference mappings.

7 Conclusions

This work presented CMOM-RS, a novel approach to CMOM predicated on the principle that, in 1:n mappings, the source entity should be semantically the disjoint union of the entities in the target expression. CMOM-RS combines two recursive strategies for candidate mapping generation that approximate this criterion: a strategy based on LLMs that exploits the semantic space conveyed by embeddings through geometric operations; and a lexical strategy that enforces disjoint union at the syntactical level (which in ontologies of related domains typically translates well to the semantic level).

Our reference-based evaluation showed that the lexical strategy is, as expected, more precise than the LLM-based strategy, as the matching syntax is a stricter requirement than the matching semantics conveyed by the embeddings space. Despite being less precise, the LLM-based strategy was able to find fully correct mappings in comparable number, while achieving near-complete coverage of the source ontologies, whereas the lexical strategy was severely restricted in coverage. These results demonstrate the effectiveness of our LLM candidate generation algorithm, validating our premise that the geometric properties of the embeddings space can be harnessed to find compositional semantic identity recursively.

Overall, our CMOM-RS approach produced the highest coverage and semantic recall in all tasks and the highest F-measure in two of the three tasks, demonstrating the value in combining the lexical and LLM strategies. Assuming the results of the reference-based evaluation can be extrapolated to the full source ontologies (rather than restricted to sources in the reference alignment), CMOM-RS can be used to expand the logical definitions currently available in these ontologies to cover nearly all their classes, greatly reducing the workload for ontologists. This is further reinforced by the results of our manual expert-based evaluation, which demonstrated the accuracy of the mappings produced by CMOM-RS for source classes not in the reference alignments.

An additional contribution of our work is the novel CMOM-RS mappings classified as correct by all experts, which we will suggest as new logical definitions to the community. Moreover, we envision that the partial reference alignments we generated and the semantic evaluation metrics we proposed for evaluating CMOM can become benchmarks to help propel research in this challenging task.

Acknowledgements

This work was supported by FCT through the fellowship 2022.11895.BD, and the LASIGE Research Unit, ref. UIDB/00408/2020 (https://doi.org/10.54499/UIDB/00408/2020) and ref. UIDP/00408/2020 (https://doi.org/10.54499/UIDP/00408/2020). It was also partially supported by the KATY project which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101017453, and by project 41, HfPT: Health from Portugal, funded by the Portuguese Plano de Recuperação e Resiliência.

References

[1] M. Abd Nikooie Pour, A. Algergawy, P. Buche, L. Castro, J. Chen, A. Coulet, J. Cufi, H. Dong, O. Fallatah, D. Faria, I. Fundulaki, S. Hertling, Y. He, I. Horrocks, M. Huschka, L. Ibanescu, S. Jain, E. Jimenez-Ruiz, N. Karam, P. Lambrix, H. Li, Y. Li, P. Monnin, E. Nasr, H. Paulheim, C. Pesquita, T. Saveta, P. Shvaiko, G. Sousa, C. Trojahn, J. Vatascinova, M. Wu, B. Yaman, O. Zamazal, and L. Zhou. Results of the ontology alignment evaluation initiative 2023. In *Proceedings of the 18th International Workshop on Ontology Matching - OM 2023*, volume 3591 of *CEUR Workshop Proceedings*, pages 97–139, Athens, Greece, 2023.

[2] S. A. Aleksander, J. Balhoff, S. Carbon, J. M. Cherry, H. J. Drabkin, D. Ebert, M. Feuermann, P. Gaudet, N. L. Harris, et al. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.

[3] A. D. Diehl, T. F. Meehan, Y. M. Bradford, M. H. Brush, W. M. Dahdul, D. S. Dougall, Y. He, D. Osumi-Sutherland, A. Ruttenberg, S. Sarntivijai, et al. The cell ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of biomedical semantics*, 7:1–10, 2016.

[4] Z. Dragisic, V. Ivanova, P. Lambrix, D. Faria, E. Jiménez-Ruiz, and C. Pesquita. User validation in ontology alignment. In *The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part I 15*, pages 200–217. Springer, 2016.

[5] M. Ehrig and J. Euzenat. Relaxed precision and recall for ontology matching. In *Proc. K-Cap 2005 workshop on Integrating ontology*, pages 25–32. No commercial editor., 2005.

[6] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2nd edition, 2013.

[7] D. Faria, E. Santos, B. S. Balasubramani, M. C. Silva, F. M. Couto, and C. Pesquita. Agreementmakerlight. *Semantic Web*, (Preprint):1–13, 2023.

[8] D. Faria, M. Silva, P. Cotovio, L. Ferraz, L. Balbi, and C. Pesquita. Results for matcha and matcha-dl in oaei 2023. 2023.

[9] G. V. Gkoutos, P. N. Schofield, and R. Hoehndorf. The anatomy of phenotype ontologies: principles, properties and applications. *Briefings in Bioinformatics*, 19(5):1008–1021, 2018.

[10] F. Gosselin and A. Zouaq. Sorbetmatcher results for oaei 2023. 2023.

[11] T. Gruetze, C. Böhm, and F. Naumann. Holistic and scalable ontology alignment for linked open data. *LDOW*, 937:1–10, 2012.

[12] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, and C. Steinbeck. Chebi in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, 44(D1):D1214–D1219, 2016.

[13] Y. He, J. Chen, D. Antonyrajah, and I. Horrocks. Bertmap: A bert-based ontology alignment system. 12 2021.

[14] S. Hertling and H. Paulheim. Order matters: matching multiple knowledge graphs. In *Proceedings of the 11th Knowledge Capture Conference*, pages 113–120, 2021.

[15] S. Hertling and H. Paulheim. Olala: Ontology matching with large language models. In *Proceedings of the 12th Knowledge Capture Conference 2023*, pages 131–139, 2023.

[16] E. Jiménez-Ruiz and B. Cuenca Grau. Logmap: Logic-based and scalable ontology matching. In *The Semantic Web–ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I 10*, pages 273–288. Springer, 2011.

[17] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[18] R. Y. Lee and P. W. Sternberg. Building a cell and anatomy ontology of caenorhabditis elegans. *Comparative and functional genomics*, 4(1): 121–126, 2003.

[19] I. Megdiche, O. Teste, and C. Trojahn. An extensible linear approach for holistic ontology matching. In *The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part I 15*, pages 393–410. Springer, 2016.

[20] C. J. Mungall, M. Bada, T. Z. Berardini, J. Deegan, A. Ireland, M. A. Harris, D. P. Hill, and J. Lomax. Cross-product extensions of the gene ontology. *Journal of biomedical informatics*, 44(1):80–86, 2011.

[21] C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis, and M. A. Haendel. Uberon, an integrative multi-species anatomy ontology. *Genome biology*, 13:1–20, 2012.

[22] D. Oliveira and C. Pesquita. Improving the interoperability of biomedical ontologies with compound alignments. *Journal of biomedical semantics*, 9(1):1–13, 2018.

[23] H. Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508, 2017.

[24] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. 2019. doi: 10.18653/v1/d19-1410.

[25] D. Ritze, C. Meilicke, O. Svab-Zamazal, and H. Stuckenschmidt. A pattern-based ontology matching approach for detecting complex correspondences. In *ISWC workshop on ontology matching, chantilly (VA US)*, pages 25–36, 2009.

[26] K. Saleem, Z. Bellahsene, and E. Hunt. Porsche: Performance oriented schema mediation. *Information Systems*, 33(7-8):637–657, 2008.

[27] M. C. Silva, D. Faria, and C. Pesquita. Matching multiple ontologies to build a knowledge graph for personalized medicine. In *European Semantic Web Conference*, pages 461–477. Springer, 2022.

[28] E. Thiéblin, O. Haemmerlé, N. Hernandez, and C. Trojahn. Survey on complex ontology matching. *Semantic Web*, 11(4):689–727, 2020.

[29] Z. Wang. Amd results for oaei 2023. 2023.

[30] L. Zhou, M. Cheatham, A. Krisnadhi, and P. Hitzler. A complex alignment benchmark: Geolink dataset. In *The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II 17*, pages 273–288. Springer, 2018.

[31] L. Zhou, M. Cheatham, and P. Hitzler. Towards association rule-based complex ontology alignment. In *Joint International Semantic Technology Conference*, pages 287–303. Springer, 2019.