Aligning Biomedical Metadata with Ontologies Using Clustering and Embeddings

Rafael S. Gonçalves(✉), Maulik R. Kamdar, and Mark A. Musen

Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA
{rafael.goncalves,maulikrk,musen}@stanford.edu

Abstract. The metadata about scientific experiments published in online repositories have been shown to suffer from a high degree of representational heterogeneity—there are often many ways to represent the same type of information, such as a *geographical location* via its latitude and longitude. To harness the potential that metadata have for discovering scientific data, it is crucial that they be represented in a uniform way that can be queried effectively. One step toward uniformly-represented metadata is to normalize the multiple, distinct field names used in metadata (e.g., *lat lon*, *lat and long*) to describe the same type of value. To that end, we present a new method based on clustering and embeddings (i.e., vector representations of words) to align metadata field names with ontology terms. We apply our method to biomedical metadata by generating embeddings for terms in biomedical ontologies from the BioPortal repository. We carried out a comparative study between our method and the NCBO Annotator, which revealed that our method yields more and substantially better alignments between metadata and ontology terms.

Keywords: Biomedical metadata · Ontologies · Alignment · Embeddings

1 Introduction

Metadata are crucial artifacts to facilitate the discovery and reuse of data. Publishers and funding agencies require scientists to share research data along with metadata. Sadly, the quality of metadata published in online repositories is typically poor. In previous work, we identified numerous anomalies in metadata hosted in the U.S. National Center for Biotechnology Information (NCBI) BioSample metadata repository [1], which describe biological samples used in biomedical experiments [5]. For example, simple binary or numeric fields are often populated with inadequate values of different data types, and there are typically many ways to represent the same information in metadata fields (e.g., *lat lon*, *Lat-Long*, among dozens of other field names to represent a location via its latitude and longitude). This is not an idiosyncratic problem affecting just a

P. Hitzler et al. (Eds.): ESWC 2019, LNCS 11503, pp. 146–161, 2019.
https://doi.org/10.1007/978-3-030-21348-0_10

single database—it is a pervasive issue that affects nearly all data repositories, and that is detrimental to the ability to effectively search and reuse data.

To make data Findable, Accessible, Interoperable, and Reusable (FAIR) [22], it is necessary that the metadata that describe the data be of high quality. One step toward generating better quality metadata out of legacy metadata is to normalize the multiple field names that are used interchangeably in metadata to describe the same thing. By doing so, we can reduce the representational heterogeneity that currently hinders metadata repositories and thus improve the searchability of the associated data. In our study of the quality of BioSample metadata, we identified that metadata authors tend to use off-the-cusp field names to annotate their datasets even when there are standardized terms (specified by the repositories) to use for a particular type of value. Metadata field names are rarely linked to terms from ontologies or metadata vocabularies. While the use of ontology terms is encouraged in some fields, there is no mechanism to enforce or verify this suggestion. Ideally, the field names used in a metadata repository should be drawn from ontologies, and there should be one single field name, rather than many, to describe each distinct type of metadata value.

In this paper, we present a new method to semi-automatically align arbitrary strings with ontology terms using *embeddings*. Embeddings are representations of words or phrases (or other types of entities) as low-dimensional numeric vectors. Word embeddings are capable of capturing contextual information and relations between different words in a text corpus [4,14]. Our method works by clustering input strings according to a string distance metric, such as the Levenshtein edit distance, and then comparing embeddings for the input strings with embeddings computed for the human-readable labels of ontology terms. The term alignments are ranked by taking into account the cosine distance between embeddings, and the edit distance between input strings and ontology term labels. To show the efficacy of our method, we align a corpus of metadata field names taken from the BioSample metadata repository with terms from ontologies in BioPortal [12]. We compare three clustering methods over six different distance metrics to determine which combination works best for the BioSample corpus. Then, we carry out a comparative study of the alignments found by our method with those found by the NCBO Annotator [7]. Finally, we conduct semi-structured interviews with subject-matter experts to verify: (a) the quality of our alignments compared to those that the NCBO Annotator finds, and (b) the appropriateness of our alignments when the NCBO Annotator finds none.

2 Related Work

The NCBO Annotator is a reference service for annotating biomedical data and text with terms from biomedical ontologies. The service works directly with ontologies hosted in the BioPortal ontology repository. The NCBO Annotator relies on the Mgrep concept recognizer, developed at the University of Michigan, to match arbitrary text against a set of dictionary terms provided by BioPortal ontologies [17]. The NCBO Annotator additionally exploits is-a relations between terms to expand the annotations found with Mgrep and to rank the alignments.

There are several recent methods to generate embeddings for entities in an RDF graph or OWL ontology. These embeddings, often called knowledge graph embeddings, require triples of the form ⟨*head entity, relation, tail entity*⟩. Translation-based methods represent each *head* or *tail* entity as a point in a vector space, and the *relation* represents a vector translation function in a hyperplane (i.e., the *head* entity vector can be translated to the *tail* entity vector, using a geometrical function over the *relation* vector) [10,19,21]. Other methods to generate these knowledge graph embeddings are inspired by language modeling approaches, which rely on sequences of words in a text corpus—that is, random walks are carried out on an RDF graph to generate sequences of entities in nearby proximity, and these sequences can be used to generate latent numerical representations of entities [16]. While these methods have been effective for link prediction in knowledge graphs, triple classification, and fact extraction, they cannot be used to assess semantic similarity of different terms by examining how they are used in the context of literature.

Word embeddings can be generated from a corpus of textual documents through popular approaches such as skip-gram and continuous bag of words models [4,14]. Word embeddings generated from a corpus of biomedical abstracts have been shown to aid in the discovery of novel drug–reaction relations [15]. Embeddings can be computed for ontology term labels by using a weighted mean of word embeddings contained within the label. Such term embeddings have been used to align elements in the vocabularies of biomedical RDF graphs [8]. Such an approach does not require re-training the term embeddings (a shortcoming of phrase embedding approaches [13]), it can use domain-specific context (e.g., the biomedical significance behind the term *tumor region*, and similarity to other terms such as *cell region* or *site of cancer*), and it can enable similarity between terms with arbitrary word lengths.

The existing works to align biomedical text with ontology terms have the following shortcomings: they do not use vector space embeddings of terms [6]; they generate ontology term embeddings using the structure of the ontology rather than background knowledge from textual data [18]; and they perform alignments against only a few ontologies containing a small number of terms. Conventional syntactic similarity metrics (e.g., Levenshtein distance) do not capture the context where terms are used. Some techniques use embeddings computed with background knowledge from Wikipedia or other generic corpora. Such embeddings are not as appropriate for our case, because biomedical metadata field names consist of multi-word, domain-specific, and complex terms, which are not typically defined or mentioned together in generic corpora [20]. Using biomedical background knowledge is essential to improve the efficacy of alignments, since most terms are biomedicine-specific; and to provide textual contexts where the aligned terms are used, which is essential to help experts verify alignments.

3 Methods

The approach we designed to align a collection of arbitrary strings with ontology terms, shown in Fig. 1, consists of the following steps. (1) We cluster input

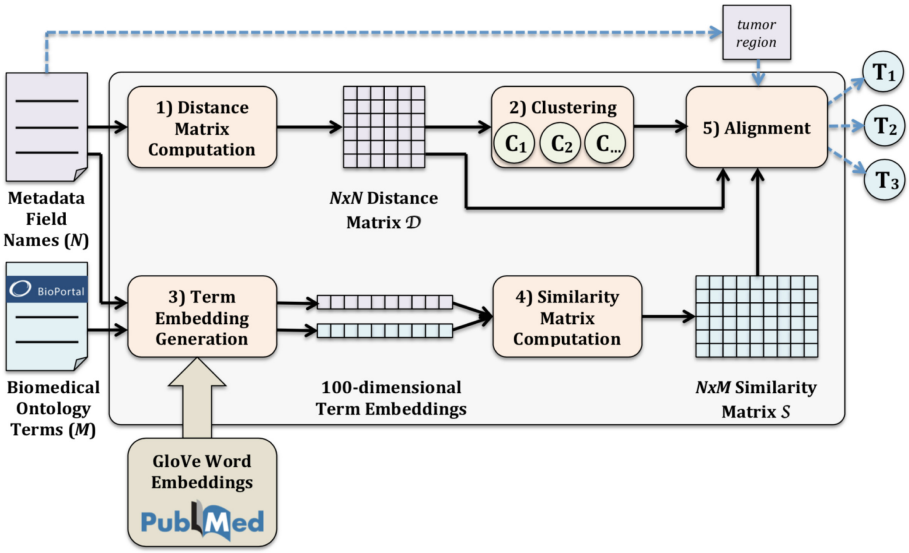

Fig. 1. Components of our metadata alignment method. In this figure, methods are colored in orange, metadata-related variables are colored in purple, and ontology-related variables are colored in blue. **(1) Distance Matrix Computation:** Our method uses a string distance metric to compare N metadata field names with each other, generating a distance matrix $\mathcal{D}$ of shape $N \times N$. **(2) Clustering:** A clustering function takes as input the distance matrix $\mathcal{D}$ and generates clusters of metadata field names. **(3) Term Embedding Generation:** Metadata field names and terms from biomedical ontologies are represented in a 100-dimensional vector space, using word embeddings generated from PubMed (https://www.ncbi.nlm.nih.gov/pubmed) abstracts. **(4) Similarity Matrix Computation:** Our method computes cosine similarity scores between term embeddings of N metadata field names and M ontology terms, generating an $N \times M$ similarity matrix $\mathcal{S}$. **(5) Alignment:** Given a metadata field name (e.g., *tumor region*), our method selects the cluster containing the field name, and then chooses the top metadata field names that have a minimal distance (using $\mathcal{D}$) from the given metadata field name. Using these highly similar field names, the distance matrix $\mathcal{D}$, and the similarity matrix $\mathcal{S}$, our method generates and ranks ontology alignments $\{T_1, T_2, T_3, \ldots\}$, shown on the right side of the figure as output. (Color figure online)

strings according to a distance metric (e.g., Levenshtein distance), which provides a grouping of syntactically similar strings. (2) We compare embeddings of input strings with ontology term embeddings trained on a corpus of biomedical ontologies. (3) We generate alignments between input strings and ontology terms by attending to the distance between the strings and the ontology terms.

For the purpose of human analysis, an ideal combination of a clustering function and a distance metric would yield relatively small clusters, for example, no bigger than 100 elements. We can test this automatically, by attending to observations such as the average size of the clusters, the size of the biggest

cluster, the cluster size variance, and so on. For the purpose of providing a reliable basis upon which to compute alignments, an ideal combination would generate clusters whose elements are so syntactically or semantically close to one another that they are likely to represent related aspects of the data. This is non-trivial to verify in any automated setting. We propose a proxy for measuring how good a set of clusters is for an alignment task such as ours, which can be automatically tested. Given an alignment function, we determine what is the maximum number of terms that can be aligned with a single ontology per cluster. Having at least one alignment for each cluster gives us a candidate alignment that we can use to inform alignments for other elements in the same cluster.

(1) Distance Matrix Computation. There are various string distance metrics shown to be useful in practice. For example, Levenshtein edit distance is widely deployed in word processors to detect typos. In our selection of distance metrics, we included the Levenshtein distance metric along with its variant, Damerau-Levenshtein. We tested cosine distance as it is the basis for our semantic alignment. We also tested Jaro distance, its variant Jaro-Winkler, and Jaccard distance (set-based, with exact match), because they are well-known and widely used metrics. We use a distance metric to compare N metadata field names with each other, thus generating a distance matrix $\mathcal{D}$ of shape $N \times N$.

(2) Clustering. Clustering functions applicable to our method must not require an upfront specification of the number of clusters. This is intuitively necessary since we have no a priori knowledge of what the data are. When selecting clustering functions, we had prior knowledge that affinity propagation [3] worked well to cluster biomedical terms according to Levenshtein distance (see [5]). We compared affinity propagation with a highly common density-based algorithm–Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [2], and with Hierarchical DBSCAN (HDBSCAN) [11], which is a hierarchical type of clustering that builds on DBSCAN. Each clustering function takes as input a distance matrix and assigns metadata field names to disjoint clusters.

(3) Term Embedding Generation. We describe the method we designed to represent metadata field names and ontology terms as high-dimensional numerical vectors, henceforth called *term embeddings*, based on the GloVe (Global Vectors for Word Representation) algorithm [14]. We generated word embeddings using a corpus of approximately 30 million PubMed abstracts from the MEDLINE database–a database of journal citations and abstracts for biomedical literature [9]. For this work, we used only the title and the abstract of each citation. We performed the same preprocessing steps to process biomedical publication abstracts (i.e., tokenization, medical entity normalization) from the PubMed repository as Percha et al. [15].

Word embeddings represent each word as a high-dimensional numerical vector based on co-occurrence counts of those words as observed in publications.

We used the GloVe algorithm to generate 100-dimensional word embeddings from the tokenized biomedical publications. The vectors are generated after a training phase of 20 iterations with an α learning rate of 0.75. These parameters (e.g., vector dimensions) are inspired by Percha et al. [15], who successfully used and evaluated biomedical word embeddings to discover novel drug–drug interactions.

We represent each metadata field name or ontology term in a high dimensional space using word embedding vectors. We generate a vocabulary of 2,531,989 words from 30 million PubMed biomedical publication abstracts. The words in our vocabulary appear in at least 5 distinct abstracts. We determined this threshold empirically, by identifying words that might be important for metadata field names (e.g., an unusual term such as `zymosterone`, which is a compound involved in the synthesis of cholesterol) and that are mentioned in only 5 abstracts. Subsequently, we represent each word in a 100-dimensional numerical space called the word embedding vector. We generate a term embedding vector by computing a weighted average of the words in the term label, with the weights being the Inverse-Document-Frequency (IDF) statistic for each word. We created a default embedding vector and IDF statistic (0.01) for any word not in the vocabulary. We show the equation to generate a term embedding vector below.

$$\mathbf{x}_{(term)} = \frac{\sum_{w_i \in \mathcal{L}(term)} idf_{(w_i)} * \mathbf{x}_{(w_i)}}{\sum_{w_i \in \mathcal{L}(term)} idf_{(w_i)}} \tag{1}$$

Here, $\mathbf{x}_{(w_i)}$ represents the 100-dimensional word embedding vector, and $\mathcal{L}(term)$ is the term label. For ontology terms, the term label $\mathcal{L}(term)$ is extracted from the values of annotations properties commonly used in biomedical ontologies to encode human-readable labels (i.e. *rdfs:label* and *skos:prefLabel*), whereas for metadata field names, the term label is the field name.

(4) Similarity Matrix Computation. We compute a cosine similarity matrix $\mathcal{S}$ for the N metadata field names and M ontology terms. Hence, the shape of matrix $\mathcal{S}$ is $N \times M$, and an individual cell in this matrix $\mathcal{S}_{ij}$ is the cosine similarity score $co\text{-}sim(N_i, M_j)$ between the metadata field name N_i and the ontology term M_j. We present the equation to compute $co\text{-}sim(N_i, M_j)$ below.

$$co\text{-}sim(N_i, M_j) = \frac{\mathbf{x}_{(N_i)} \cdot \mathbf{x}_{(M_j)}}{||\mathbf{x}_{(N_i)}||||\mathbf{x}_{(M_j)}||} = \frac{\sum_{k=1}^{n=100} x_{(N_i)k} * x_{(M_j)k}}{\sqrt{\sum_{k=1}^{n=100} x^2_{(N_i)k}} \sqrt{\sum_{k=1}^{n=100} x^2_{(M_j)k}}} \tag{2}$$

In the above equation, $\mathbf{x}_{(N_i)}$ and $\mathbf{x}_{(M_j)}$ are the 100-dimensional term embedding vectors for the metadata field name N_i and the ontology term M_j respectively.

(5) Alignment. We specify our approach to find alignments between strings and ontology terms in Definition 1.

Definition 1. *Let* align$_r$ (S, C$_\Psi$, D$_m$) *be a function that takes a set of strings* S*, a collection of clusters* C *computed according to a clustering function* Ψ*, and*

a matrix D *of the pairwise distances between all strings in* S *computed according to a distance metric* m*, and returns an alignment map* $\mathcal{A} := \mathrm{S} \rightarrow \mathcal{T}$ *that maps each input string in* S *to a list of recommended ontology terms* $\mathcal{T}$*. We say a string* $s \in S$ *is aligned with an ontology term* $t \in \mathcal{T}$ *if the average of the cosine similarity between the embeddings of* s *and* t*, denoted* $w(s)$ *and* $w(t)$ *respectively, and the edit similarity between* s *and* t *is above a threshold* r*, as follows.*

$$\mathcal{A} := \{s \rightarrow t \mid s \in S \wedge \frac{\text{co-sim}(w(s), w(t)) + \text{edit-sim}(s, t)}{2} > r\} \tag{3}$$

We use both the cosine and edit similarities between strings to prevent false positive alignments between completely unrelated strings, which could result from aligning using cosine similarity alone. We have anecdotal evidence that the IDF ranking sometimes unduly attributes too much weight to certain words in metadata field names and throws off alignments. On the other hand, by taking both measures into account we run the danger of missing alignments between strings that are syntactically very different but semantically equivalent. We experimented with different weighting schemes between cosine and edit distance, and found the best compromise to be the average between *co-sim* and *edit-sim*.

Materials. The metadata in the NCBI BioSample are specified through fields that take the form of *name–value pairs*, for example, **geo location:** *Alaska*. Users of BioSample can use standard *field names* specified by BioSample, or they can (and often do) coin new names for their fields. We gathered all the fields names used across all metadata in BioSample. There are a total of 18,198 syntactically distinct field names in the BioSample metadata. We then normalized the metadata field names by replacing all non-alphabetic symbols with spaces, splitting words concatenated in CamelCase notation, converting all strings to lower case, and trimming all but one space between words. After removing duplicates, and strings which had fewer than 3 characters that we considered to be inappropriately short as metadata field names, we ended up with 15,553 unique field names. To generate embeddings, we used a corpus of ontologies from BioPortal composed of 675 ontologies, which was extracted on May 25, 2018.

4 Results

In this section we present the results of clustering and aligning the metadata field names from the BioSample repository with terms from ontologies in the BioPortal repository. We first carried out an exploratory study of clustering functions and distance metrics in order to identify a suitable combination for the BioSample corpus. Based on this experiment, described in Sect. 4.1, we hoped to draw a theory about selecting an ideal combination of a clustering function and distance metric that can be automated in the future. Subsequently, we performed

a comparative study of the alignments found by our method and those found by the NCBO Annotator. We compare the two methods according to whether they found at least one alignment per cluster, and according to how many field names they were able to align. We describe this study in Sect. 4.2.

4.1 Clustering Metadata Field Names

In this experiment, we normalized the BioSample field names as described in Sect. 3, and then we evaluated combinations of clustering methods and distance metrics according to the following measures: number of clusters generated; average, median, and standard deviation of the number of field names in each cluster; number of field names in the biggest and smallest cluster; and number of clusters of the smallest size. We show the results in Table 1.

The HDBSCAN algorithm generated clusters that were of reasonable size for human analysis. The average cluster size across the board was around 20 field names, and the median cluster size was 10 or fewer. However, every distance metric we used resulted in at least one huge cluster with over 1,500 field names, and hundreds of tiny clusters with only 2 names each. Overall the clusters generated using HDBSCAN were not particularly suitable for human analysis.

Table 1. Results of clustering metadata field names. The table shows, from left to right, for each combination of clustering function and distance metric: the number of clusters; the average cluster size; the median cluster size; the standard deviation of the cluster size; the size of the biggest cluster; the size of the smallest cluster; and the number of clusters of the smallest size. The best results are shown in bold.

Clustering method	Distance metric	Nr. of clusters	Avg size	Median size	Std dev.	Biggest cluster	Smallest cluster	Nr. clusters of smallest size
HDBSCAN	Levenshtein	641	27	10	94	1671	2	140
	Damerau	712	24	8	85	1671	2	174
	Jaro	871	20	7	91	2232	2	230
	Jaro-Winkler	1053	16	6	80	2385	2	255
	Jaccard	721	24	9	81	1762	2	178
	Cosine	1381	12	6	89	3091	2	314
DBSCAN	Levenshtein	1191	14	2	279	8099	2	805
	Damerau	1178	14	2	280	8037	2	795
	Jaro	1623	10	2	192	7006	2	924
	Jaro-Winkler	941	18	2	345	10151	2	514
	Jaccard	1045	16	2	306	7660	2	755
	Cosine	1150	15	2	258	6844	2	702
AP	Levenshtein	2145	8	5	8	60	1	491
	Damerau	2146	8	5	8	58	1	495
	Jaro	1495	11	10	7	50	1	1
	Jaro-Winkler	**1387**	**12**	**10**	**9**	**61**	**2**	**13**
	Jaccard	1743	10	8	7	43	1	264
	Cosine	1415	12	7	32	1126	1	159

The results of clustering metadata field names using DBSCAN were very poor with all distance metrics. This method yielded extremely large clusters containing around 8,000 field names and over 500 clusters with only 2 field names. There is significant variability in the size of the clusters too, as illustrated by the high standard deviation of the size of clusters. Generally, the clusters generated using DBSCAN are even less suitable for human analysis than HDBSCAN.

Clustering metadata field names using affinity propagation resulted in seemingly high quality clusters in terms of size. There were on average 1,783 clusters across all distance metrics, which is higher than the other clustering methods. However, this means that the clusters are generally much smaller. On average there were 10 field names per cluster, and the median size was 8 field names. The sizes of the large clusters found by affinity propagation were especially encouraging; the largest cluster was found using the Jaro-Winkler distance metric and it has 61 field names. In comparison with the other methods, this is highly encouraging. The number of clusters of the smallest size varies between distance metrics. Clustering based on the Levenshtein, Damerau, or Jaccard distances results in many clusters with only one field name. Using the Jaro-Winkler distance as a basis for clustering seems to yield the best overall results—it results in a desirable number of metadata field names per cluster, it has low variability of cluster size, and it results in neither too big clusters nor too small clusters.

4.2 Alignments Between Field Names and Ontology Terms

In this experiment, we aimed to verify how well our approach can align clusters and their elements with ontologies and ontology terms, using different combinations of clustering methods and distance metrics. We compare the alignments obtained using our approach with those obtained using the NCBO Annotator. For each cluster in a set of clusters, we set out to find one ontology that provides the most alignments for the elements in that cluster—that is, an *ontology recommendation* for a cluster—and we record the number of field names aligned with that ontology. The alignments of our method are based on a minimum cosine similarity (threshold r in Definition 1) of 0.85 between the embeddings of field names and ontology terms. The results of this experiment are in Table 2.

Our method found an ontology with which to align cluster elements for over 90% of clusters, on average, across all combinations of clustering methods and distance metrics. On the other hand, the NCBO Annotator only found ontologies for less than 50% of clusters on average. The best coverage obtained using the NCBO Annotator is using affinity propagation and Jaro-Winkler distance, for which 69% of the clusters were aligned with ontologies. While this combination is not the best that our method found, if we take into account the average coverage of both our method and the NCBO Annotator, the combination just mentioned is still the best one, yielding 81% average coverage for both methods.

Using our method, we aligned 7 metadata field names per cluster with ontology terms across all combinations in Table 2, while the NCBO Annotator only aligned 2 field names, on average. Taking into account that the mean cluster size is 12 elements using affinity propagation and Jaro-Winkler, aligning 8 of those

Table 2. Comparison of alignments found by our approach and by the NCBO Annotator. The table shows for each combination of clustering method and distance metric: the number of ontology recommendations obtained for all clusters ("Nr. recs."); the percentage of clusters that have one ontology recommendation, i.e., coverage ("Cov."); the average/median number of field names in each cluster that were aligned with ontology terms. The best results are shown in bold.

Clustering method	Distance metric	Our approach			NCBO annotator		
		Nr. recs.	Cov.	Avg./median fields covered	Nr. recs.	Cov.	Avg./median fields covered
HDBSCAN	Levenshtein	625	98%	11/5	390	61%	3/1
	Damerau	695	98%	10/4	428	60%	3/2
	Jaro	853	98%	9/4	403	46%	3/2
	Jaro-Winkler	1027	98%	8/4	678	64%	2/1
	Jaccard	711	99%	10/4	441	61%	3/1
	Cosine	1355	98%	6/4	859	62%	2/1
DBSCAN	Levenshtein	1042	87%	7/2	309	26%	1/1
	Damerau	1031	88%	7/2	306	26%	2/1
	Jaro	1448	89%	6/2	642	40%	2/1
	Jaro-Winkler	821	87%	9/2	358	38%	2/1
	Jaccard	937	90%	7/2	375	36%	1/1
	Cosine	1037	90%	7/2	440	38%	1/1
AP	Levenshtein	1962	91%	5/4	773	36%	3/1
	Damerau	1961	91%	5/4	774	36%	3/1
	Jaro	1432	96%	7/6	867	58%	2/2
	Jaro-Winkler	**1295**	**93%**	**8/7**	**952**	**69%**	**2/2**
	Jaccard	1609	92%	6/5	1066	61%	2/2
	Cosine	1356	96%	7/5	913	65%	2/2

elements on average is a success. The most metadata field names aligned per cluster was achieved using the HDBSCAN algorithm. This is unsurprising, since HDBSCAN produces much larger clusters, and so it is more likely that it can find many more alignments per cluster and thus have better overall coverage.

Overall, using the NCBO Annotator we found alignments for 12,454 metadata field names, while using our method we found alignments for all terms. The average similarity score of the topmost alignments for all field names is 0.94 (where 1 means an exact match), which looks highly promising.

5 Evaluation by Expert Panel

To investigate the quality of the alignments found by our method, we designed an evaluation with a panel of subject-matter experts to verify the following two hypotheses. *(H1)* The ontology term alignments provided by our method are preferable to alignments provided by the NCBO Annotator for the same metadata field name. *(H2)* The ontology term alignments provided by our method are adequate to describe metadata field names even when the NCBO Annotator

does not provide alignments for those field names. We continued to use a minimum similarity score (threshold r in Definition 1) of 0.85 between BioSample metadata field name and ontology terms, to ensure that our method provided good quality alignments.

5.1 Semi-structured Interview Design

We designed a semi-structured interview format to test our hypotheses, and we conducted interviews with four experts who are biomedical metadata experts. Our panel experts were specifically selected as they all have extensive experience in: engineering and working with metadata authoring and publishing software; constructing metadata forms for specific datasets and community standards; using ontologies and metadata vocabularies to annotate data; and working with users and developers to build infrastructure for describing scientific data. All experts have backgrounds and degrees in computer science.

We compiled one list of metadata field names to test each hypothesis. *List 1* to test *H1* was composed of 6 metadata field names randomly drawn from the set of metadata field names for which both our method and the NCBO Annotator found alignments. *List 2* to test *H2* was composed of 6 metadata field names randomly drawn from the set of metadata field names for which the NCBO Annotator did not find alignments. The selected metadata field names for the expert evaluation are listed in Table 3. The questions for each expert were randomly selected from List 1 and 2 in such a way that, for every metadata field name, we would have 3 responses from the experts.

Table 3. Lists of metadata field names used in the interviews. The table shows the metadata field names selected for *List 1* to test *H1*, and for *List 2* to test *H2*.

List 1		List 2	
Q1	sample depth m	Q7	scientific name
Q2	mapped reference genome	Q8	patientnumber
Q3	tissue source	Q9	participants
Q4	isolate id	Q10	stimuli
Q5	tumor region	Q11	cryptophytes
Q6	cardiovascular failure	Q12	relhumidityavg

In our study we asked the experts to discuss if and how each metadata field name is related to the ontology terms with which it is aligned. We use qualitative categories, which we describe in Table 4, to gather feedback from experts. We chose this style of scale since a typical Likert-style scale is unlikely to be informative as to precisely how a given term relates to a field name (according to the experts' understanding and interpretation of the text). For example, it is preferable to know, for a given metadata field name *bird species*, that the

ontology term *species* is **more general** than *bird species*, or that *owl* is **more specific**, than it is to know that an ontology term is more or less appropriate to describe, or better or worse aligned with some field name.

Table 4. Expert response categories used in the interviews. The table shows the categories that we used to gather feedback about how metadata field names related to the recommended ontology terms. In the second column we show example recommended ontology terms for the input field name *"bird species"*.

Category	Examples for *"bird species"*
Unrelated	Mouse
Looks unrelated	Bird by species
May be related	Bird species identifier
More general	Species
More specific	Owl
Looks similar	Bird species name
Identical	Species of bird

The format of the interview was as follows:

1. We described the purpose of the study—to evaluate different algorithms for aligning metadata field names with ontology terms.
2. We described the challenge that our method meets—to normalize the many different field names used in metadata records to describe the same data.
3. We listed typical questions that experts could ask of the interviewer—to give example values for the metadata fields; to clarify the task or the questions; to clarify or give examples of the categories in Table 4.
4. We described the task that the experts were about to carry out—for each question out of 9 questions:
   (a) We showed the metadata field name, and asked the expert to describe her/his understanding of the field name. We ranked their understanding according to whether the meaning of the field name was *clear*, *roughly clear*, or *unclear*.
   (b) We showed a list of human-readable labels from the ontology terms that were aligned with the field name.
   (c) We asked the expert to discuss each ontology term along the categories described in Table 4 while voicing their reasoning.

5.2 Expert Panel Results

The results of testing *H1*, shown in Fig. 2, suggest that many of the alignments found by our approach were of very high quality—in the *identical* category. On the other hand, none of the alignments found by the NCBO Annotator

were considered identical to the metadata field names. Among the alignments found by our method, none was considered *unrelated* to the corresponding field name, while the NCBO Annotator suggested 12 such *unrelated* alignments. For example, the NCBO Annotator suggested the term *mapped* for the field name *mapped reference genome* (Q2).

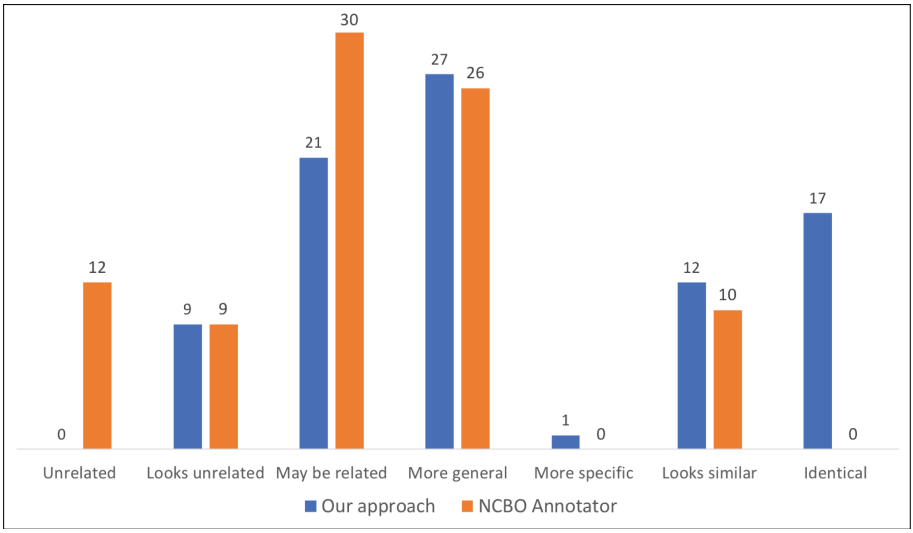

Fig. 2. Results of expert panel. Number of alignments that were considered to be in each of the categories along the x-axis, for both our method and the NCBO Annotator.

Both methods yielded 9 alignments that *looked unrelated* to the field names. One example of these is the ontology term *cardiovascular function* for the field name *cardiovascular failure* (Q6), as categorized by one expert. On the other hand, this expert considered the ontology term *cardiovascular* (given by the NCBO Annotator) to be *similar* to *cardiovascular failure*. One school of thought may consider *cardiovascular failure* to be *more specific* than *cardiovascular function*. In this case, the expert reasoning could also be that *cardiovascular function* indicates "heart is functioning well" and *cardiovascular failure* indicates "heart is not functioning well". In the future, we will show experts a small snippet of all descendants of the aligned ontology term (with definition) for more context.

The alignments found by the NCBO Annotator were often categorized as *more general* than the metadata field names. For example, for the field name *tissue source* (Q3) the NCBO Annotator suggested *tissue* and *source*, while our method suggested *tissue source site* and *source organ*. As expected, we found that when the metadata field names were syntactically identical or very similar to the labels of ontology terms, experts categorized them as *identical*. For example, experts categorized *depth of sample* as either identical or similar to the field name

sample depth m. One expert observed that *depth of sample* is more generic, since it is missing the specification of the unit of measurement.

Interestingly, when answering Q8, two experts thought the meaning of the metadata field name *patientnumber* was clear—they saw the first ontology term recommended, which happened to be "patient number" and they said it was identical. When they saw the third term recommended, which was "number of patients", both experts revised their interpretation of the meaning of the field name. One expert lowered the alignment of both options to "similar", while the other expert admitted that the third option could be just as "identical" as the first option to align with the field name.

The results of testing *H2* demonstrate that most of the alignments found by our method were very precise, and categorized as either *similar* or *identical* by participants. The study showed that 39 (43%) ontology term alignments were considered *identical* matches to the metadata field names; 23 (26%) alignments were considered *similar*; 7 (8%) alignments *more specific* than the field name; 10 (11%) alignments *more general*; 8 alignments that *may be related*; 1 alignment that *looked unrelated*; and 2 alignments that were considered *unrelated*.

6 Conclusions

We developed a new method based on clustering and embeddings to align arbitrary strings with ontology terms. We applied our method to align metadata field names from NCBI BioSample metadata with terms from ontologies in the BioPortal repository. In our experiments we determined that clustering metadata field names using affinity propagation according to the Jaro-Winkler distance metric was the most suitable combination for our corpus. Using this combination as the basis of our alignment method, we were able to find high quality alignments between all metadata field names and at least one ontology term. Unlike existing ontology alignment or string similarity methods, we compute semantic similarity using a combination of background knowledge derived from biomedical publication abstracts and ontology term descriptions, and we can efficiently align strings against a corpus of 10 million terms from 675 ontologies.

We carried out a comparative study between our method and the NCBO Annotator, in which we discovered that our method found many more alignments overall. We also compared the two methods as part of an expert panel that we conducted via semi-structured interviews. We discovered that our method was able to find highly precise alignments even between strings that, while syntactically not the same, described the same thing. On the other hand, none of the alignments found by the NCBO Annotator were considered identical to the metadata field names. The overall results of our expert panel are illustrative of the efficacy of our method, and its potential for applications. In this work we tuned and applied our method to biomedical metadata, although in principle our approach can be applied to any domain that has a sufficiently rich source of textual knowledge and ontologies to generate embeddings from.

Our experiments show that our method is a suitable solution to align biomedical metadata with ontology terms. Aligning and replacing legacy field names in

a metadata repository (or multiple repositories) with ontology terms can substantially improve the searchability of the metadata, and thus the discoverability of the associated data. The applications of our method are twofold: it can be used in metadata authoring software to give real-time suggestions for metadata field names, or to give suggestions for field values that should correspond to ontology terms; and it can be used to facilitate metadata cleaning, by equipping scientists with a means to align metadata field names with ontologies.

In future work, we will implement a generic metadata field recommendation service, and subsequently integrate it with the CEDAR metadata authoring tool for recommending (a) field names when users are building metadata templates, and (b) field values when users are filling in metadata templates. We will design a standalone tool that uses our method to: provide a visualization of clusters formed out of input metadata field names; recommend mappings between the field names and ontology terms; give users ranked options for the best terms for specific fields; and generate a new enriched dataset that materializes the selected field name mappings by replacing legacy field names with ontology terms.

Acknowledgments. This work is supported by grant U54 AI117925 awarded by the U.S. National Institute of Allergy and Infectious Diseases (NIAID) through funds provided by the Big Data to Knowledge (BD2K) initiative. BioPortal has been supported by the NIH Common Fund under grant U54 HG004028.

We thank the experts in our evaluation panel: John Graybeal, Josef Hardi, Marcos Martínez-Romero, and Csongor Nyulas (all of whom from the Center for Biomedical Informatics Research at Stanford University), for their participation.

References

1. Barrett, T., et al.: BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucl. Acids Res. **40**, D57–D63 (2012)
2. Ester, M., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Conference on Knowledge Discovery and Data Mining (1996)
3. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**(5814), 972–976 (2007)
4. Goldberg, Y., Levy, O.: Word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722 (2014)
5. Gonçalves, R.S., Musen, M.A.: The variable quality of metadata about biological samples used in biomedical experiments. Sci. Data **6**, 190021 (2018)
6. Jiménez-Ruiz, E., Cuenca Grau, B.: LogMap: logic-based and scalable ontology matching. In: Aroyo, L., et al. (eds.) ISWC 2011. LNCS, vol. 7031, pp. 273–288. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25073-6_18
7. Jonquet, C., et al.: NCBO annotator: semantic annotation of biomedical data. In: International Semantic Web Conference (2009)
8. Kamdar, M.R., et al.: An empirical meta-analysis of the life sciences (linked?) open data cloud (2018). http://onto-apps.stanford.edu/lslodminer
9. Koster, C., Seutter, M., Seibert, O.: Parsing the medline corpus. In: Recent Advances in Natural Language Processing (2007)

10. Lin, Y., et al.: Learning entity and relation embeddings for knowledge graph completion. In: AAAI Conference on Artificial Intelligence (2015)
11. McInnes, L., Healy, J., Astels, S.: HDBSCAN: hierarchical density based clustering. J. Open Source Softw. **2**(11), 205 (2017)
12. Noy, N.F., et al.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucl. Acids Res. **37**, W170–W173 (2009)
13. Passos, A., Kumar, V., McCallum, A.: Lexicon infused phrase embeddings for named entity resolution. arXiv preprint arXiv:1404.5367 (2014)
14. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (2014)
15. Percha, B., Altman, R.B., Wren, J.: A global network of biomedical relationships derived from text. Bioinformatics **1**, 11 (2018)
16. Ristoski, P., Paulheim, H.: RDF2Vec: RDF graph embeddings for data mining. In: Groth, P., et al. (eds.) ISWC 2016. LNCS, vol. 9981, pp. 498–514. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46523-4_30
17. Shah, N.H., et al.: Comparison of concept recognizers for building the open biomedical annotator. In: BMC Bioinformatics, vol. 10, p. S14. BioMed Central (2009)
18. Smaili, F.Z., Gao, X., Hoehndorf, R.: OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. arXiv preprint arXiv:1804.10922 (2018)
19. Socher, R., et al.: Reasoning with neural tensor networks for knowledge base completion. In: Advances in Neural Information Processing Systems (2013)
20. Wang, Y., et al.: A comparison of word embeddings for the biomedical natural language processing. J. Biomed. Inform. **87**, 12–20 (2018)
21. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph embedding by translating on hyperplanes. In: AAAI Conference on Artificial Intelligence (2014)
22. Wilkinson, M.D., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data **3**, 160018 (2016)