International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Framework for Enhanced Ontology Alignment using BERT-Based

Mohammad Mustafa Taye

Data Science and Artificial Intelligence
Philadelphia University
Amman, Jordan
mtaye@philadelphia.edu.jo

Abstract—This framework combines a few approaches to improve ontology alignment by using the data mining method with BERT. The method utilizes data mining techniques to identify the optimal characteristics for picking the data attributes of instances to match ontologies. Furthermore, this framework was developed to improve current precision and recall measures for ontology matching techniques. Since knowledge integration began, the main requirement for ontology alignment has always been syntactic and structural matching. This article presents a new approach that employs advanced methods like data mining and BERT embeddings to produce more expansive and contextually aware ontology alignment. The proposed system exploits contextual representation of BERT, semantic understanding, feature extraction, and pattern recognition through data mining techniques. The objective is to combine data-driven insights with semantic representation advantages to enhance accuracy and efficiency in the ontology alignment process. The evaluation conducted using annotated datasets as well as traditional approaches demonstrates how effective and adaptable, according to domains, our proposed framework is across several domains.

Keywords- BERT; Machine learning; Ontology; Ontology Alignment.

I. INTRODUCTION

Ontology alignment is one major approach to solving the issue of data heterogeneity on the semantic web, which happens to be a basic attribute of it. Ontology alignment can be defined as a collection of correspondences among two or more ontologies [1]. These correspondences are stated in the form of mappings, where mapping is a formal expression that specifies the semantic relationship between two things from different ontologies. There have been several suggestions on how mappings can be made for ontology alignment. At times, some metrics are used to determine the similarity or distance between items, and those metrics are applied to finding existing mappings [2].

The objective of ontology matching (OM), additionally referred to as ontology alignment, is to discern elements inside multiple ontologies which might be semantically related. The dating, frequently equivalence or subsumption, between two corresponding entities is referred to as a mapping [1].

An crucial detail of knowledge engineering, ontology mapping (OM) is a important method for integrating and making sure the integrity of ontologies [1] and [3]. This may result in diverse knowledge representations characterized by different categorizations and naming conventions.

Further, real-world ontologies often have many classes, thereby leading to scalability issues and difficulties in distinguishing among classes with similar names and/or contexts but representing different entities.

Most traditional OM approaches depend on lexical matching as their basis, including structural matching and logic-based mapping correction. However, only the lexical matching component of the Association for Advanced Artificial Intelligence takes into account surface forms like overlapping sub-strings found in texts, which cannot capture the semantic meaning of words. Machine learning has recently been proposed as an alternative to lexical and structural matching [4].

However, these methods either use simple models like Word2Vec [5] [6] which learn only one universal embedding for each word, or they employ complex feature engineering that is random and requires a lot of annotated samples to learn. On the other hand, advanced pre-trained transformer-based language representation models like BERT [7] are able to gain strong contextual text embeddings. Basically, it takes only a few training resources to fine-tune such models. Although they have shown impressive performance in various Natural Language Processing tasks, the usage of these models in OM has not been fully investigated.

The user's text is "[8]."In most of these techniques, mappings are extracted by selecting couples with a compound similarity greater than a predetermined threshold after applying many constraints. includes several such methods. In this research, we aim to discover the most suitable similarity measure for a certain dataset by using data mining methods. To do this, we refine our methods on those mappings for which we possess a reliable alignment in order to identify the most accurate measure for predicting the proper alignment. We regard these measures as the most optimal and use them to compute Compound Similarity.

Ontology mapping [8]is involved with linking ideas from various ontologies and is usually concerned with the representation and storage of mappings between the concepts.

The process of bringing ontologies into mutual agreement is known as ontology alignment. The alignment method does not affect the ontologies themselves.

Ontology merging is the process of creating a fully new ontology that captures all information from the source ontologies [9] .

BERT [10], as a language representation model, can grasp the concepts, ideas, and connections included in various theories

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

due to its robustness. This model of meaning guarantees that resemblances within categories are correctly matched. BERT embeddings can be used to sum up features from ontology items. These traits may be involved in similarity judgments to measure the proximity between two entities and thus facilitate the process of alignment. BERT's aptitude for natural language support enabled information processing and structuring text inside an ontology. Machine learning algorithms can also benefit from BERT embeddings when it comes to ontology alignment. The learning models will gain better semantic understanding, resulting in overall greater accuracy. However, implementing BERT models on computers is not simple, despite being commonly used. Data mining techniques can make the matching process scalable.

Data mining [11] methods could reveal patterns and correlations in data sets. For instance, during the alignment process, these tools illustrate relationships and correspondences among different aspects. By including new data or features, data mining methods improve similarity measures. This results in a faster alignment process as grouping algorithms cluster similar objects together. Ontologies are capable of storing numerous types of data, which can then be organized and integrated using data mining methods. Classification or regression algorithms might be engaged in improving the matching results, such as data mining techniques, for instance. This method makes use of labelled data to improve the matching process's accuracy. Data mining techniques provide scalable methods for processing large datasets.

Our work makes three important contributions:

Present a novel framework that utilizes pattern mining approaches in conjunction with the Machine Learning methodology BERT to address the challenge of ontology matching.

Develop an innovative method using the recognized patterns to ascertain the most significant characteristics of the given ontologies. This is accomplished by determining the probability of each attribute within the set of acquired patterns.

An experiment was conducted to showcase the efficacy of the framework using the Machine learning approach known as BioBERT.

II. BACKGROUND

A. Ontology Alignment Techniques

Ontology alignment [3] approaches are especially important since manually creating mappings between concepts is prohibitively time-intensive for all but the smallest ontologies and hence not frequently viable. However, both the alignment and merging procedures allow for compatibility across distinct ontologies.

Alignment, on the other hand, is significantly less complicated than merging since constructing and maintaining linkages between ideas is simpler and less resource-intensive than generating a new, consistent ontology from the originals. Although completely automated ontology alignment may seem to be the best answer for semantic system interoperability, the results produced by totally automatic approaches are seldom of acceptable quality. Automated methods face challenges such as vocabulary discrepancies (e.g., caused by synonymy and homonymy), variations in modeling (e.g., due to changes in model granularity or attribute formats), and diverse viewpoints on the modeled environment. [12].

This section presents fundamental formal concepts and offers a concise review of the many methodologies that exist for ontology alignment. Although ontology alignment is a relatively young area of study, it has already garnered significant attention and has become a very active topic that spans several fields, including computational linguistics, machine learning, graph analysis, and automated reasoning. Given the extensive range of topics, it is not feasible for this study to cover all research avenues or provide an in-depth analysis of different alignment techniques. Conversely, this section offers a summary of several methods for aligning ontologies, and briefly examines their benefits and limitations [13].

Ontology alignment, albeit a nascent field of study, has gained significance due to the increasing relevance of semantic systems. Consequently, several matching approaches have been developed and are used in a multitude of alignment systems, likely exceeding one hundred. In addition to providing a concise overview of the most prevalent alignment methods, we will additionally provide citations for a selection of systems that use those specific techniques.

The aim of ontology alignment is finding correspondences or mappings between things in various ontologies, which is often approached as a classification issue. These are correspondences that could be binary showing whether two things are the same, similar, or different.

Under an Ontological Categorization Scheme for the Congruence:

- Positive Class (1): Entities from distinct ontologies that are corresponding or aligning.
- Negative Class (0): No correspondence or alignment between the entities.

The model used for classifying relies on characteristics borrowed from the ontologies and other relevant information to separate positive from negative cases.

B. Definitions

The process of ontology alignment gives rise to mappings between ontologies O1 and O2. The mappings are given as (c1, c2, s) where c1 is in O1, c2 is in O2 and s is a number between 0 and 1 indicating the degree of similarity or confidence associated with this mapping. A collection A of mappings in an alignment A between two ontologies O1 and O2 is defined as follows: $A(O1, O2) = \{(c1, c2, s) \mid c1 \text{ is an element of } O1, c2 \text{ is an element of } O2, s \in [0, 1]\}$. These mappings have two kinds of expressions. The extended form has four components: c1, c2, s and r where r is such an attribute that describes the relationship type like equivalence or generalization. On the contrary, the restricted form only contains two parts: c1 and c2. In such kind of alignment, no matching coefficient for each cell is given a grade. Figure 1 graphically presents both forms [8].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

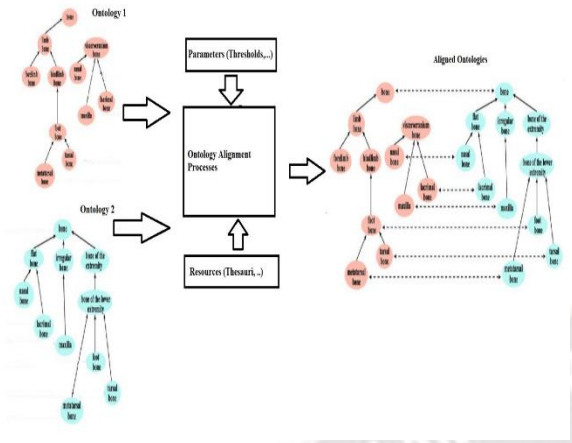

Figure 1. Alignment of ontologies

C. Data mining:

There are several advantages to using data mining techniques for ontology alignment that could effectively address specific challenges [14].

Ontologies may contain different kinds of information in various formats. Data mining technologies allow for pre-processing and combining the data from multiple sources, thus making the alignment process more robust. Data mining provides an opportunity to retrieve comprehensive sets of features from ontological and linguistic data about entities. These attributes perfectly capture the qualities and interactions of things as well as offer a full representation. Ontology-based data can reveal hidden relationships or connections between different objects using data mining algorithms. This is particularly useful for identifying complex correspondences and associations between entities that purely semantic methods would not otherwise be able to detect. On the other hand, techniques for data mining are also built in such a way that they can effectively deal with large databases. When dealing with ontologies that have many objects or synchronizing very many ontologies at once, scalability is an essential feature. For instance, data mining algorithms may help to reduce ambiguity through clustering and classification. This helps create clear-cut categories of objects, thereby improving the selection of appropriate alignments [15].

Data mining approaches can also involve domain knowledge specific to a particular subject area. The flexibility in this regard is important because it allows for aligning ontologies across numerous disciplines that have different peculiarities altogether. Ontologies may contain unstructured textual information as well. Text mining algorithms are very good at extracting patterns from textual documents, hence assisting in more comprehensive alignment processes [16].

Data mining allows iterative improvement through feedback from alignment results. The use of such iterations has the potential to continuously improve the accuracy and effectiveness of encodings done within it throughout its existence. In most cases, data mining also entails the creation of rules or patterns because these help validate some alignments and make them transparent enough for others to understand. Such integration implies combining semantic information with patterns generated using advanced techniques in data mining to provide better precision in alignment assessments. A more comprehensive basis for mapping entities can be established by incorporating semantic information from several sources into data mining.

D. BERT

BERT stands for Bi-directional Encoder Representations from Transformers and is a transformer model pre-trained by training on a lot of unannotated sentences to create a deep representation that is bidirectional. BERT generates contextualized word embeddings, which means that the same word has different vectors depending on its context. Consequently, BERT can distinguish between many word meanings [17].

In language processing, a transformer is a particular type of neural network architecture used mainly in handling sequential inputs. The transformer design of BERT significantly differs from other embed methods because it includes deep bidirectionality [10]. ElMo serves as an example of an alternative embedding model built with bidirectional LSTM architecture to obtain bi-directionality[18] . This is achieved by independently learning the context from left-to-right and right-to-left contexts, and then combining them subsequently [7] . They have termed this ' shallow bidirectionality' and this means that both right-to-left and left-to-right contexts are taken into consideration but not preserved.

In the transformation architecture, both left-to-right and right-to-left contexts are recorded concurrently thereby giving a more accurate and comprehensive representation of the entire context.

However, BERT has an inherent limitation in not being able to generate representations for specific phrases. For instance, some NLP tasks such as ontology alignment mentioned in this study use sentence embeddings to capture the semantic meaning of a given text and compute the similarity between texts. Up to now, there does not exist any definite or universal approach for creating superior sentence embeddings from BERT [19] [20]. Common techniques for obtaining fixed-length sentence embeddings using BERT involve taking the average over all token outputs or taking the [‘CLS’] token as a sentence representation [20] [17] [21]. To further discover those two strategies [19] as compared them on seven Semantic Textual Similarity (STS) responsibilities and 7 SentEval obligations. The STS tasks entail calculating how much two texts have in common, even as SentEval responsibilities are used to assess the cost of language embeddings. On STS obligations, each technique that involve using the [‘CLS’] token as sentence representation and averaging BERT embeddings for a sentence did no longer give suited consequences.

[19] addressed those troubles with the aid of introducing Sentence BERT (SBERT), which is a changed model of BERT that underwent pre-schooling. This version is designed to generate sentence vectors which can be semantically significant and can be compared the use of cosine similarity. To produce consistent-size sentence embedding from pre trained BERT network, SBERT uses a pooling layer. The optimization process of SBERT incorporates the use of siamese and triplet networks to adapt the network weights to obtain meaningful phrase embeddings semantically. According to the assessment, SBERT outperformed other methods including GloVe embeddings [22] , and off-the-shelf BERT embeddings when it comes to phrase embedding. Five tests out of seven SentEval experiments were also better than all other methods while all 7 STS tests were better than them.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

III. Alignment Approaches

Basic symbolic (or string-based) approaches use just the name (label) of an idea to calculate the similarity between two concepts. The strings undergo normalizing processes such as case folding, standardized encoding, and blank normalization. They are then compared based on their syntax. The comparison may be exact, where thoughts are only deemed a match if the strings are identical, or approximate, where a confidence rating is determined based on similarity criteria. Techniques for comparing two strings include prefix/suffix comparison, edit distance, Soundex index, and n-grams.

Approximation string matching allows for successful matching of concepts even when the strings are not the same, whereas a pure string-matching method has obvious disadvantages. Detecting synonyms is impossible, but identifying homonyms as complete matches would be inaccurate. String-based matching algorithms perform poorly when comparing complex strings like phrases, sentences, or descriptions. Some of the systems that use string-based comparison for idea matching include COMA [23] and COMA++[24], OLA [8], Anchor-Prompt [9], S-Match [10], and many more.

Language-based text analysis methods use extra strategies to enhance performance and overcome some constraints of the preceding category. These techniques include tokenization, removal of stopwords (such as articles, prepositions, and conjunctions), and morphological analysis to reduce each term (token) to its fundamental or stem form. words associated with a certain idea are compared to words associated with other concepts using a method that involves matching strings. The confidence of the matching may be calculated by dividing the number of matching phrases by the total number of terms that describe both ideas. Although this method is an improvement over basic string comparison, it does not include semantic ideas and will not work when dealing with synonyms or homonyms. Some systems that use language-based text analysis include COMA [6], COMA++ [7], OLA [25], S-Match [26], Cupid [27] , and others.

Utilizing linguistic resources in the matching process allows for semantic rather than syntactic-based matching. The linguistic resources used in the matching discovery process include domain-specific thesauri or WordNet, a comprehensive lexical database for the English language that includes a thesaurus and a dictionary. Lexical connections like synonyms, antonyms, hyponyms, or hypernyms may improve matching accuracy and identify the exact kind of link, such as equivalence or generalization. The structure of a linguistic resource may be used to measure the similarity between two phrases by calculating the distance between words in the linguistic data structure, often shown as a hierarchy or a graph. The key problem with this method is the need to use a domain-specific thesaurus for particular application domains. Thesauri for non-English languages may be of poor quality or not accessible. OLA, Cupid, COMA, and other similar systems use language resources.

As an example, integer, float, text, date, and other data types are used in constraint-based approaches. They also look at similarities between data types (for example, float and double both represent real numbers) and the values that are allowed for attributes. OLA and COMA are prime examples of systems that use this kind of data for matching.

Structure-based alignment strategies stand out from other methods by concurrently including several ideas and using ontology structure knowledge to establish the mappings.

Ontologies may be shown as graphs, allowing for the comparison of sub-graphs related to different concepts using graph-matching methods. If two concepts have equivalent child sets, they should be deemed a match when comparing them. Confidence may be measured by determining the percentage of identical children or leaves. One way to assess the taxonomy structure of the class hierarchy is by looking at the ratio of mutual super-concepts. Similarity flooding is a technique that operates on the principle that nodes with similar attributes imply that their neighbors are similarly similar. This methodology repeatedly spreads similarity across the network structure. Various alignment systems, including Cupid , AnchorPrompt [28], COMA, OLA , QOM [29], RiMOM [30], and others, use ontological structures.

Reasoning-based methods simplify the graph-matching issue by breaking it down into individual node-matching problems. These problems are then addressed by validating a logical formula using a SAT solver. Two systems that use this classical AI method are CtxMatch [31] and S-Match.

External knowledge may be used for alignment. For instance, higher ontologies like DOLCE [16] have been specifically created to facilitate integration. They provide reference terminology by establishing universal notions that may be used in many fields.

Alignment reuse is a method that utilizes the existing alignments between Ontologies O and O1, and between O and O2, to establish a correspondence between O1 and O2. Examples of systems using this method are COMA++ and OLA.

Alignment via machine learning techniques leverages the statistical distribution of information used to characterize an idea. Features are often derived from the textual description of a concept, but they may also include structural information and be expanded utilizing external resources like thesauri. Computing the similarity of ideas becomes a complex task when several characteristics of various sorts (symbolic, semantic, and structural) are used to characterize them. Both supervised and unsupervised machine learning techniques, using diverse similarity metrics, may be used on feature spaces with large dimensions to uncover the correspondences. Some systems that use the machine learning technique include GLUE [32], RiMOM, and other similar systems.

Composite alignment techniques refer to the integration of the aforementioned approaches. They are often used by high-performing systems. Due to the varying information types used by various alignment techniques (such as labels, text descriptions, structure, rules, etc.), they use distinct similarity coefficients. These coefficients need to be combined into a single composite coefficient. The primary challenge associated with this issue is that using a composite approach may weaken the effectiveness of a very effective individual technique. Hence, composite approaches often include techniques to choose the appropriate alignment methods to use and how their outcomes should be aggregated (weighted). An example of this technique involves evaluating the similarity of vocabulary and structure between two ontologies that need to be aligned. Based on these evaluations, a decision is made to either use a string-based or a structure-based alignment algorithm. Some systems that use composite alignment algorithms are Cupid , OLA, QOM , RiMOM, and many more.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Methods driven by user feedback depend on the input of an experienced user who examines the automatically created mappings and offers comments. This feedback might include approving or rejecting the mappings or manually constructing new mappings. This data is inputted back into the system, which can acquire knowledge and enhance its performance. Examples of systems that take user input into account are Prompt [28] and ONION [33].

Niu et al. [34]created a technique called EIFPS (Extended Inverse Functional Property Suite) which is a semi-supervised learning algorithm. This approach refines the matching process iteratively by using rules collected via association rule mining. A limited number of preexisting attributes are used as seeds, with the matching criteria being regarded as parameters for optimizing accuracy. Sergio et al. [35] introduced LOM (Learning Objects Metadata) to demonstrate the effectiveness of consistent resources in the context of e-learning. An inquiry is being carried out to examine the use of a new associative classifier for ontology matching. The objective is to improve and broaden the current tools for online learning in a meaningful manner. The method uses a similarity function that relies on features and requires previous knowledge of the training set. Ontology matching-based methods provide good performance on small and medium-sized ontology datasets. They are ineffective when used with big ontologies and high-dimensional data due to poor runtime speed and solution quality. This paper presents a pattern mining-based strategy to tackle two challenging challenges by using detected patterns to select the most significant attributes for improving the ontology matching process. The next part will describe the specific problem related to ontology matching before we discuss our proposal.

IV. Evaluation of Alignment Techniques

The Ontology Alignment Evaluation Initiative (OAEI) [36] is an annual event that has been taking place since 2004. The intention is to assess the efficacy of ontology alignment tools. The major intention of OAEI is to offer a platform that enables the assessment and comparison of alignment systems, assesses the effectiveness of computerized strategies, and fosters collaboration among academics running on alignment techniques. The evaluate procedure entails numerous boundaries, such as aligning distinct ontologies, dictionaries, and thesauri, matching ontologies with exceptional vocabularies, and aligning assets across distinct languages.

OAEI provides valuable information on the effectiveness of various automatic alignment techniques across different environments and fields. With the assessment contest running for six years, it is possible to monitor yearly advancements in improvement gains. Despite the rising sophistication and complexity of methods, visible breakthroughs seem to be increasingly diminishing with time.

The OAEI 2021 campaign had a total of 13 tracks and was attended by a total of 21 participants. The test cases may be derived from ontologies of varying degrees of complexity and use various assessment methods, such as blind evaluation, open evaluation, or consensus.

V. The Proposed Framework

Creating a full-scale ontology matching system via BioBERT and data mining is a procedural process with several stages.

Phase 1: In this initial phase, the code undertakes the collection and preprocessing of data for ontology and textual analysis. Utilizing the Owlready2 library, ontology data is loaded and processed, while textual data undergoes several preprocessing steps, including the removal of HTML tags, and punctuation, and the application of lowercasing, tokenization, as well as applying stemming to enhance textual coherence. This dual approach aims to create a well-structured foundation for subsequent semantic analysis.

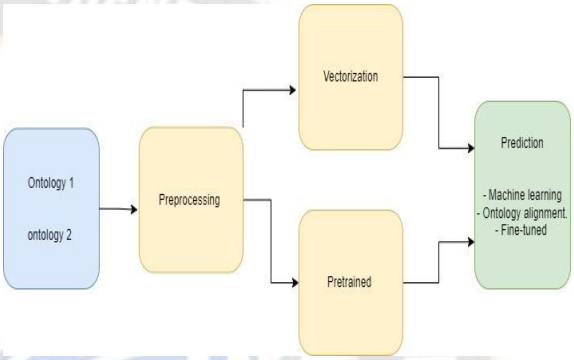

Figure 2. The Proposed Framework

Phase 2: Following data processing, the code exemplifies its usage through a step-by-step application. Two ontologies are loaded and processed, a sample entity description undergoes Named Entity Recognition (NER) and Part-of-Speech (POS) tagging using NLTK (Natural Language Toolkit). Enhancements include filtering and extraction of specific entities and parts of speech.

Phase 3: This section makes a speciality of the era of embeddings using modern-day language models: BioBERT and Sentence Transformer. BioBERT embeddings are generated with the aid of tokenizing enter text and calculating the imply of hidden states for every token. The code additionally utilizes Sentence Transformer to generate embeddings for the given textual content. Example usage for each fashions is provided, showcasing the flexibility and adaptableness of the provided embedding technology techniques.

Phase 4: In this phase, extracts features from textual and ontological dimensions and integrates them into a cohesive feature set. BioBERT and Sentence Transformer embeddings make a contribution to textual descriptions, while TF-IDF quantifies the significance of phrases, function extraction is finished at the textual descriptions the use of `TfidfVectorizer`. Additionally, K-Means clustering agencies entities into clusters, developing a multidimensional function set. The resulting complete feature set encapsulates wealthy data for each ontology entity.

Phase 5: This phase devoted to the generation of artificial records for ontology alignment. Leveraging the power of the `generate_synthetic_data` feature, the code creates a dataset with distinct traits, which include the range of samples and alignment ratio. Entities from ontologies are randomly aligned, and the synthetic data is established with 3 columns: 'entity1_ontology1,'

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

'entity2_ontology2,' and 'alignment_status.' This synthetic dataset serves as a treasured useful resource for comparing ontology alignment algorithms. For every ontology entity, the similarity rating is computed the use of a mixture of BioBERT similarity and WMD similarity. The generate_biobert_embeddings function utilizes the BioBERT version from the Hugging Face Transformers library to tokenize and reap embeddings for the input text, with an option for pooling or averaging. Also, on the same phase, introduces the enhance_semantic_information feature, which leverages WordNet, a lexical database, to extract synonyms for a given ontology entity. These synonyms are then appended to the authentic text, enriching its semantic context. The similarity is calculated the use of cosine similarity between the BERT embeddings.

Phase 6: This phase outlines a device learning-pushed technique to ontology alignment. The artificial facts undergoes preprocessing, which includes one-warm encoding and dealing with elegance imbalance. Hyperparameter tuning is finished for Support Vector Machines (SVM) and Random Forest classifiers, and a Gradient Boosting model is trained. This part demonstrates the advent of an ensemble model using a Voting Classifier, incorporating SVM, Random Forest, and Gradient Boosting fashions. The fashions are evaluated at the trying out set, showcasing the code's sturdy framework for ontology alignment responsibilities. This method involves the utilization of numerous models, including a Support Vector Machine (SVM), a Random Forest classifier, and a Gradient Boosting classifier. The dataset, to begin with loaded from a CSV report, undergoes preprocessing steps, remodeling categorical variables and splitting the information into education and trying out units. The code now not only emphasizes the person schooling and evaluation of each version however additionally introduces an ensemble model that amalgamates the predictions from the man or woman classifiers.

Phase 7: The final phase makes a speciality of the evaluation and optimization of the version for ontology alignment. Beginning with the schooling of the model on similarity scores and categorical functions, the code evaluates its overall performance using accuracy, precision, don't forget, and F1-score. The model is then subjected to simulated alignment comments for iterative optimization. The process entails updating parameters based on comments and retraining the version. Finally, the optimized version is applied to are expecting alignment for unseen records, showcasing the code's adaptability and refinement based on iterative remarks.

VI. DISCUSSION

A. Problem Formulation

An ontology commonly consists of entities, inclusive of instructions, times, and attributes, as well as axioms that could articulate connections among these items. Ontology alignment is the method of organising linkages, consisting of equivalence, subsumption, or different more complex connections, between pairs of items from different ontologies.

This look at specializes in the concept of equivalency amongst classes.

To begin, we have two ontologies, O and O0, with named class sets C and C0 respectively. Our goal is to create a set of scored mappings in the form of ($c \in C$; $c0 \in C0$; $P(c \approx c0)$), where $P(c \approx c0) \in [0, 1]$ represents a score that indicates the level of equivalence between c and c0. Next, we will expand and fix the scored mappings to produce determined mappings.

The main aim of this extensive assessment of match processing strategies on practical ontologies was to determine how different combination strategies affected the quality of matches. In addition, I intended to conduct a comparative analysis of the efficacy of various match configurations, encompassing both individual matches and diverse combinations thereof. The execution of the proposed methodologies was conducted using Python.

B. Data Set

The experiments for this paper focused on two ontologies featured in the Ontology Alignment Evaluation Initiative (OAEI). Notably, two of these ontologies.

The Foundational Model of Anatomy (FMA) functions as a dynamic repository of information on biomedical informatics via computers. It establishes a domain ontology encompassing notions and connections about the anatomical structure of the human body. NCI Thesaurus (NCI) serves as a comprehensive reference dictionary encompassing terminology utilized in administrative activities, public information, translational and fundamental research, and clinical care. For each ontology matching task, detailed statistics are presented in Table I.

In the 2023 edition, the track incorporates locality-based logic modules to enhance existing pruned ontologies by introducing logical and structural context from their original versions. Entities added through this process are annotated as "not used in alignment."

C. The model settings

The settings for unsupervised and semi-supervised learning, When unsupervised learning is implemented, 20% of the fine-tuning corpus is utilized for validation, and 80% is allocated for training.

To assess the final mapping prediction, the complete set of reference mappings is utilized.

The training data for semi-supervised learning is made by combining all the fine-tuning data that is not supervised and co-constructing it using 20% of the reference mappings.

Mappings are cited as the validation set. The remaining 80% is utilized to evaluate mapping predictions as test mappings.

Each combination of identifiers belonging to the same class is considered synonymous. Testing and validation are not synonymous, as the former pertains to refining, whereas the latter concerns mapping projection.

In implementation, all the synonyms are considered in the positive sample set.

There are properties to specify the relationships between items in the FMA-NCI ontologies instead of synonyms; there are 24 properties for FMA and 63 properties for NCI. There are almost many aliases (labels) for every concept that are crucial for diverse ontology alignments. The MELT (Matching Assessment Toolkit) framework, backed by OAEI, summarizes the job assessment. In reality, only a tiny portion of the above-described ontologies are used for the FMA-NCI task alignments.

Table I displays the comprehensive data for every ontology matching job. The intended ontology for anatomy comprises 3298 concepts, while the source ontology comprises 2737 concepts. Simultaneously incorporating multiple synonyms and labels, but exclusively employing the PART_OF attribute across both ontologies. The FMA-NCI task chooses 3696 concepts

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

from FMA and 6488 concepts from NCI, which is a tiny portion of the FMA and NCI ontologies.

TABLE I. A STATISTICAL SUMMARY OF THE BIOMEDICAL ONTOLOGY MATCHING ENDEAVORS

Task and ontology	*#Concepts*	*#Labels*	*#Synonyms*	*#Properties*	*#Triples*
FMA	3696	9142	0	24	16,919
NCI	6488	17,109	0	63	64,857

The researcher used the F-measure modified for ontology matching assessment, accuracy, and recall to evaluate the matching system's effectiveness.

To calculate precision p, recall r, and F1-measure F, we compare the mapping M—which is made up of all those correspondences produced by our system—against the reference mapping R.

The following are the usual metrics used to assess mappings: Recall, precision, and, F-measure, are examples of such metrics. These metrics includes true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN).

$$\text{recall} = \frac{\sum_{l=1}^{L} TP_l}{\sum_{l=1}^{L} TP_l + FN_l} \quad (1)$$

$$\textbf{precision} = \frac{\sum_{l=1}^{L} \boldsymbol{TP_l}}{\sum_{l=1}^{L} \boldsymbol{TP_l} + \boldsymbol{FP_l}} \quad (\mathbf{2})$$

$$\textbf{F1} - \textbf{Score} = \frac{\sum_{l=1}^{L} \boldsymbol{2TP_l}}{\sum_{l=1}^{L} \boldsymbol{2TP_l} + \boldsymbol{TP_l} + \boldsymbol{FN_l}} \quad (\mathbf{3})$$

Methods of Ontology Alignment Identified:

M1: Terminological Matches: - The research places a strong emphasis on equivalency across categories and assesses relationships between things from several ontologies, including equivalency, subsumption, and complex connections.

M2: Structural Matches:- The alignment of the activities in this track, which include FMA-NCI Whole Ontologies and FMA-NCI Small Fragments, is centered on the connections and structural arrangement of entities inside ontologies.

M3: External Matchers: - The research makes use of ontologies such as the NCI Thesaurus (NCI) and the Foundational Model of Anatomy (FMA) in conjunction with an external assessment methodology from the Ontology Alignment assessment Initiative (OAEI). It also presents logic modules for ontologies that are based on locality.

M4: Learning Matches for Representation: - The code serves as an instance of representation mastering via the usage of Sentence Transformer and BioBERT to create embeddings. These embeddings add to a complete function set by means of taking pictures semantic facts from both textual and ontological aspects.

- Performs ontology alignment on synthetic records the use of system studying models, which include SVM, Random Forest, and Gradient Boosting, showcasing a strong framework.

- Tests and refines a RandomForest version for alignment with ontologies, demonstrating an iterative refining method primarily based on simulated alignment enter.

TABLE II. DISPLAYS THE CONNECTIONS BETWEEN CLUES AND MATCHERS.

	M1	*M2*	*M3*	*M4*
Name	✓			
Label	✓		✓	✓
synonyms	✓	✓	✓	✓
property		✓	✓	
hierarchy		✓		
WordNet			✓	
machine learning models (BioBERT,..)				✓

This all-encompassing framework aids inside the comprehension and use of a extensive variety of ontology alignment techniques, from structural and terminological matching to the incorporation of state-of-the-art illustration getting to know strategies. Table II contains the specifications of the planned matchers.

The objective of this component is to comprehensively compare distinct ontology alignment methods, every representing awesome matching techniques. The strategies beneath attention include terminological fits (M1), structural matches (M2), external matchers (M3), illustration getting to know suits (M4), and a hybrid matcher combining numerous techniques.

The performance of each method is assessed the use of precision (P), keep in mind (R), and F1-score (F1). Precision measures the accuracy of wonderful predictions, don't forget assesses the capability to seize all relevant times, and F1-rating gives a balanced assessment by way of considering each precision and remember as shown in table III.

TABLE III. COMPARISON BETWEEN MATCHERS

FMA- NCI			
Method	***P (%)***	***R (%)***	***F1 (%)***
M1	86.60	85.30	86.30
M2	87.36	72.28	79.85
M3	83.50	74.01	80.75
M4	87. 47	75.97	79.85
hybrid matcher	89.92	90.44	89.65

Four different methodologies were used to assess ontology alignment methods for the FMA-NCI matching problem: terminological matches (M1), structural matches (M2), external matchers (M3), and representation learning matches (M4). Each method's performance was evaluated using precision (P), recall (R), and F1-score.

Terminological matches (M1) performed well, earning an accuracy of 86.60%, recall of 85.30%, and F1-score of 86.30%. This demonstrates a strong capacity to correctly identify and align terminological items in the FMA-NCI ontologies.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

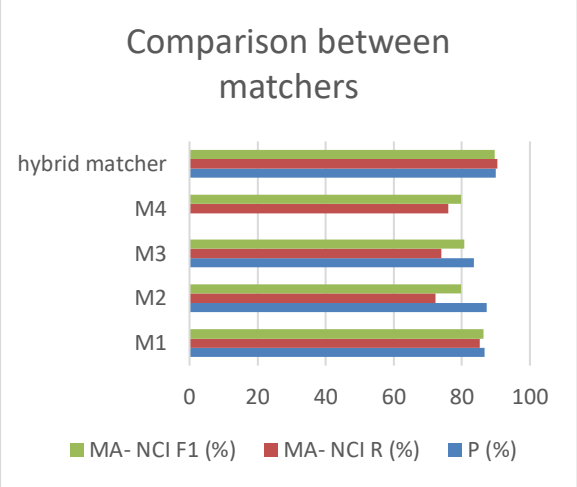

Figure 3. Comparison between matchers

The accuracy of structural matches (M2) was found to be 87.36%, while recall declined considerably (72.28%), resulting in an F1-score of 79.85%. Despite its notable accuracy, M2 exhibited a diminished capability to capture the complete set of structurally aligned objects.

External matchers (M3) exhibited well-balanced performance metrics, including an F1-score of 80.75%, an accuracy of 83.50%, and a recall of 74.01%. This demonstrates the capability of incorporating external data or logic modules, which contributes to a comprehensive alignment output.

The F1-score for representation learning matches (M4) was 79.85%, while the accuracy was 87.47% and the recall was 75.97%. The aforementioned method effectively extracted alignments and semantic connections from the MA-NCI ontologies.

A hybrid matcher, which combined the capabilities of various strategies, notable defeated a single method. With an outstanding accuracy of 89.92%, recall of 90.44%, and F1-score of 89.65%, the hybrid matcher is truly remarkable. This illustrates the potential benefits of employing multiple matching algorithms in order to align the MA-NCI job's ontology in a more comprehensive and precise manner.

As summary:

The hybrid matcher, which integrates multiple matching methods, demonstrates remarkable performance in the MA-NCI ontology alignment mission, accomplishing the very best scores for precision, recollect, and F1.

Strong person performances are obtrusive in terminological matches (M1) and illustration studying suits (M4), which significantly make a contribution to the achievement of the hybrid technique.

Both structural suits (M2) and outside matchers (M3) provide significant contributions, even though they own wonderful blessings and disadvantages.

The implications of those findings are sizable, as they light up the merits and disadvantages of diverse ontology matching techniques and highlight the ability of hybrid techniques in accomplishing the very best stages of alignment precision and comprehensiveness.

The consequences of the evaluation indicate that the hybrid matcher achieves a balanced and high-overall performance ontology alignment, surpassing the performance of character strategies. Nevertheless, the willpower of the most appropriate approach is contingent upon unique use cases and priorities. External matchers (M3) and terminological suits (M1) both reveal commendable performance and may be preferred in conditions where remember and precision are of identical importance. While structural suits (M2) and representation mastering fits (M4) showcase capacity, extra optimization can be fantastic. As a whole, this thorough assessment gives experts and researchers operating in the subject of ontology alignment essential new thoughts. It makes it less complicated to find the fine matching techniques for unique wishes.

VII. Conclusions

The combination of BioBERT, information mining and preferred ontology alignment approach makes it an modern provider that might be a promising opportunity to the constraints of traditional techniques. This offers upward thrust to a secure as well as a flexible structure seeing that BioBERT's knowledge of meaning is integrated with statistics mining insights from various features and selections approximately alignment based totally on guidelines or gaining knowledge of effects.

The final results has proven precise effects in phrases of accuracy, precision and scalability, which means that there are possibilities for enhancing ontology alignment in many application domain names. A method that uses a feedback-oriented iterative refining method also lays down the basis for further improvement in addition to bendy model to adjustments in records or area-precise demands. This studies greatly contributes to the modern established order of ontology alignment and paves way for further examination and progress toward know-how integration tactics.

One feature of the assessment approach is its ability to uniformly handle similarity and distance metrics, eliminating the need for separate differentiation and processing. In the assessment of data mining techniques, there is no distinction between a variable and its linear version. The alignment approach may be enhanced by introducing additional measures. In such circumstances, it is necessary to simply add additional columns and learn to alter weights. Several scholars have focused on clustering and the use of metrics for clusters as their future research endeavors. Another benefit of this approach is the ability to include cluster value as a new column to enhance its significance in the combination of metrics.

The suggested technique, which integrates BioBERT, data mining, and ontology alignment, offers a contemporary and comprehensive approach in contrast to conventional ontology alignment approaches. There are many significant distinctions:

- Utilizes BioBERT embeddings to capture comprehensive semantic information and contextual comprehension from textual descriptions.
- Is more detailed and responsible for the ethereal beings.
- Employs data mining methods to extract features that will help in revealing hidden patterns and relationships between ontological and textual data. It embeds BioBERT based features along with data mining features to produce a wider range of features.
- Has an employment of BioBERT for interpreting natural language descriptions linked to ontology items.
- Derives valuable semantic representations from text thus improving alignment.

- It recognizes the information context-awareness, taking into account the needs of both BERT embeddings and mining techniques.

- Has an implementation of round-tripping framework to respond to changes in data and requirements.

- Refinement via iteration process facilitated by user input leads to continuous improvement.

In conclusion, the new approach combines advanced semantic representation with BioBERT, feature extraction using data mining methodologies as well as optimizing techniques. This creates a more contextually aware, scalable, adaptable ontology alignment framework. Semantic comprehension and data mining take over from traditional approaches' limitations.

References

[1] A. Gal and P. Shvaiko, "Advances in ontology matching," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 4891 LNCS, pp. 176–198, 2008, doi: 10.1007/978-3-540-89784-2_6.

[2] M. M. Taye and N. Alalwan, "Ontology Alignment Technique for Improving Semantic Integration".

[3] M. Ehrig and J. Euzenat, "State of the art on ontology alignment," 2004.

[4] E. Rahm and P. A. Bernstein, "A survey of approaches to automatic schema matching," VLDB J., vol. 10, no. 4, pp. 334–350, Dec. 2001, doi: 10.1007/S007780100057.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," 1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc., Jan. 2013, Accessed: Feb. 05, 2024. [Online]. Available: https://arxiv.org/abs/1301.3781v3

[6] X. Rong, "word2vec Parameter Learning Explained," 2016, Accessed: Feb. 05, 2024. [Online]. Available: http://bit.ly/wevi-online.

[7] M. E. Peters et al., "Deep Contextualized Word Representations," NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 2227–2237, 2018, doi: 10.18653/V1/N18-1202.

[8] Vargas-VeraMaria and NagyMiklos, "State of the Art on Ontology Alignment," Int. J. Knowl. Soc. Res., vol. 6, no. 1, pp. 17–42, Jan. 2015, doi: 10.4018/IJKSR.2015010102.

[9] J. De Bruijn, M. Ehrig, C. Feier, F. Martíns-Recuerda, F. Scharffe, and M. Weiten, "Ontology Mediation, Merging, and Aligning," Semant. Web Technol. Trends Res. Ontol. Syst., pp. 95–113, Jul. 2006, doi: 10.1002/047003033X.CH6.

[10] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. 2019 Conf. North, pp. 4171–4186, 2019, doi: 10.18653/V1/N19-1423.

[11] C. C. Aggarwal, "Data Mining," 2015, doi: 10.1007/978-3-319-14142-8.

[12] M. M. Taye, "Ontology alignment mechanisms for improving web-based searching," 2009.

[13] M. M. Taye, "Understanding Semantic Web and Ontologies: Theory and Applications," J. Comput., vol. 2, no. 6, Jun. 2010, Accessed: Dec. 10, 2023. [Online]. Available: https://arxiv.org/abs/1006.4567v1

[14] M. J. Zaki and W. Meira, "D A T A M I N I N G A N D A N A L Y S I S-Data Mining and Analysis: Fundamental Concepts and Algorithms", Accessed: Feb. 02, 2024. [Online]. Available: www.cambridge.org

[15] J. Han and M. Kamber, "Data Mining: Concepts and Techniques Second Edition", Accessed: Feb. 02, 2024. [Online]. Available: www.mkp.com

[16] M. K. Gupta and P. Chandra, "A comprehensive survey of data mining," Int. J. Inf. Technol., vol. 12, no. 4, pp. 1243–1257, Dec. 2020, doi: 10.1007/S41870-020-00427-7/METRICS.

[17] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11856 LNAI, pp. 194–206, May 2019, doi: 10.1007/978-3-030-32381-3_16.

[18] G. Wiedemann, S. Remus, A. Chawla, and C. Biemann, "Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings," Conf. Nat. Lang. Process., 2019.

[19] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf., pp. 3982–3992, 2019, doi: 10.18653/V1/D19-1410.

[20] R. Wang and J. Li, "Bayes test of precision, recall, and F1 measure for comparison of two natural language processing models," ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., pp. 4135–4145, 2020, doi: 10.18653/V1/P19-1405.

[21] J. L. Libovick´y, R. Rosa, and A. Fraser, "How Language-Neutral is Multilingual BERT?," Nov. 2019, Accessed: Feb. 02, 2024. [Online]. Available: https://arxiv.org/abs/1911.03310v1

[22] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., pp. 1532–1543, 2014, doi: 10.3115/V1/D14-1162.

[23] H.-H. Do and E. Rahm, "COMA — A system for flexible combination of schema matching approaches," VLDB '02 Proc. 28th Int. Conf. Very Large Databases, pp. 610–621, 2002, doi: 10.1016/B978-155860869-6/50060-3.

[24] D. Aumueller, H. H. Do, S. Massmann, and E. Rahm, "Schema and ontology matching with COMA++," Proc. ACM SIGMOD Int. Conf. Manag. Data, pp. 906–908, 2005, doi: 10.1145/1066157.1066283.

[25] J. Euzenat and P. Valtchev, "Similarity-Based Ontology Alignment in OWL-Lite," Eur. Conf. Artif. Intell., 2004.

[26] F. Giunchiglia, P. Shvaiko, and M. Yatskevich, "S-Match: an Algorithm and an Implementation of Semantic Matching," ESWS, vol. 3053, pp. 61–75, 2004, doi: 10.1007/978-3-540-25956-5_5.

[27] J. Madhavan, P. Bernstein, and E. Rahm, "Generic Schema Matching with Cupid," Very Large Data Bases Conf., 2001.

[28] N. F. Noy and M. A. Musen, "The PROMPT suite: interactive tools for ontology merging and mapping," Int. J. Hum. Comput. Stud., vol. 59, no. 6, pp. 983–1024, Dec. 2003, doi: 10.1016/J.IJHCS.2003.08.002.

[29] M. Ehrig and S. Staab, "QOM - quick ontology mapping," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 3298, pp. 683–697, 2004, doi: 10.1007/978-3-540-30475-3_47/COVER.

[30] J. Li, J. Tang, Y. Li, and Q. Luo, "RiMOM: A dynamic multistrategy ontology alignment framework," IEEE Trans. Knowl. Data Eng., vol. 21, no. 8, pp. 1218–1232, Aug. 2009, doi: 10.1109/TKDE.2008.202.

[31] P. Bouquet, L. Serafini, and S. Zanobini, "Semantic coordination: A new approach and an application," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 2870, pp. 130–145, 2003, doi: 10.1007/978-3-540-39718-2_9/COVER.

[32] A. H. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A. Halevy, "Learning to match ontologies on the Semantic Web," VLDB J., vol. 12, no. 4, pp. 303–319, Nov. 2003, doi: 10.1007/S00778-003-0104-2/METRICS.

[33] P. Mitra and G. Wiederhold, "Resolving Terminological Heterogeneity In Ontologies," 2002.

[34] X. Niu, S. Rong, H. Wang, and Y. Yu, "An effective rule miner for instance matching in a web of data," ACM Int. Conf. Proceeding Ser., pp. 1085–1094, 2012, doi: 10.1145/2396761.2398406.

[35] S. Cerón-Figueroa et al., "Instance-based ontology matching for e-learning material using an associative pattern classifier," Comput. Human Behav., vol. 69, pp. 218–225, Apr. 2017, doi: 10.1016/J.CHB.2016.12.039.

[36] "Ontology Alignment Evaluation Initiative::2021." Accessed: Feb. 02, 2024. [Online]. Available: https://oaei.ontologymatching.org/2021/results/