*Artificial Intelligence Research and Development*
*T. Alsinet et al. (Eds.)*


*doi:10.3233/FAIA240452*

# Cross-Domain Linkage for the Music Industry: An Industrial and Synthetic Benchmark

Donghao ZHANG [a] Emilio MOLINA [b] Aglaia GALATA [b] Gonçal CALVO [b] Denis GUILHOT [b] Martí SÁNCHEZ-FIBLA [a,c,1]

[a] *DTIC, Universitat Pompeu Fabra, 08018, Barcelona, Spain*
[b] *BMAT Licensing S.L. Barcelona, Spain*
[c] *Artificial Intelligence Research Institute, IIIA (CSIC), Campus UAB, 08193, Spain*

**Abstract** Record linkage across diverse domains is a challenging task with industrial applications ranging from medical records to social media identity linkage. In this study, we present a complex case in the music industry: linking incomplete metadata across domains, from Sound Recordings (SRs) to their corresponding Musical Works (WKs). We present a definition of the problem and highlight its key aspects: comparing record fields beyond conventional string similarity; matching lists of names that only partially align; applying attribute rules, as some attribute values may reflect the quality of information; and applying contextual rules, since the match between an SR and a WK should be evaluated within the context consisting of related WKs. We present a synthetic benchmark that replicates the complexities of the real-world industry problems. While not the focus of the paper, we also report preliminary results of a Transformer-based model that leverages pre-trained embeddings of entity attribute values along with information from the aforementioned key aspects.

**Keywords.** Natural Language Processing, Data Fusion, Record Linkage, Entity Alignment, Entity Matching, Embeddings

## 1. Introduction

The digitalization of the music industry has led to exponential growth in content creation and consumption, with streaming platforms like Spotify indexing over 100,000 new tracks daily, and YouTube adding huge amounts of unstructured user-generated content, just to mention the mainstream ones. Music metadata matching has become challenging and particularly problematic when metadata comes from these heterogeneous sources, hindering the creation of master databases and links between musical entities. We focus here on a significant source of inefficiency within the music industry: the inability to accurately reconcile textual metadata records across the myriad databases involved in the royalty distribution process. This is one of the many tasks in which BMAT specializes.

Addressing this issue is critical for improving operational efficiency and ensuring fair and timely royalty payments. But despite the market's expansion, challenges persist due to complex systems and fragmented ownership of copyright. This has resulted in inefficiencies where artists struggle to receive fair compensation, exemplified by the

[1]Corresponding Author, E-mail: marti.sanchez@iiia.csic.es

"black box" of royalties worth over 400$ million in the US alone. One of the main causes of these inefficiencies in the music industry is the inability to determine the correspondence between records from the hundreds of databases involved in royalty distribution processes across various parts of the value chain: from consumption data on streaming platforms, through phonogram management databases, to systems managing composition catalogues to identify rights holders. Existing matching technologies fail to respond accurately when confronted with records that often contain incomplete or incorrect information.

We depart from this complex scenario with the compilation of datasets provided by BMAT that are representative of the metadata matching problems in the music industry. The record matching problem is translated into finding the correspondence of entities from two domains, in our case, Sound Recordings (SR) and their associated Musical Work (WK). SRs being the recordings of musical compositions (WKs) which can have the same or different titles and could share or not artists with the corresponding composers, (or more generally) contributors of the work (WK). This matching problem is challenging due to several factors: metadata from both domains (SRs and WKs) comes from different sources (companies, legal agencies, etc.), correspondence may depend on the complex priorities within the music industry, as well as being subject to regional rules, among other reasons. BMAT has been tackling this problem by creating datasets of matching and non-matching (SR-SR and SR-WK) records with human annotations, enabling the training of Machine Learning models for this problem.

The Matching Learning project (Red.Es, reference 2021/C005/00149157) initiated a collaboration of the research teams in UPF (and now also at IIIA, CSIC) with BMAT, to incorporate new state of the art Machine Learning techniques to solve the complex matching problem. This paper is the first outcome of a fruitful collaboration with several contributions: defining the problem in formal terms and identifying analogous problems in the state of the art (section 2), presenting the BMAT industrial in-house proprietary benchmark (section 2.1), designing a new synthetic benchmark that preserves the challenges of the industry case and could become public (section 3), present a high-level description of a Transformer-based architecture model that addresses the problem and which leverages the use of pre-trained Large Language Models (LLMs) and uses relational information (inline with Graph Neural Networks and Knowledge Graphs) presented in section 4 and finally preliminary results on both benchmarks (section 5). The technical details of the model are not intended to be presented here, only the high-level components to set an initial reference point for the new benchmark.

## 2. Problem definition

The Record Matching problem has been identified as a particular case of Record Linkage which come under different statements and naming such as Entity Resolution [2, 11] and Entity Alignment (EA) [4, 7, 13, 14]. In Entity Resolution, the goal is to uniquely identify the same entity (a record of metadata with attributes and their values) across various sources (typically databases). In Entity Alignment, the objective is to link the same entity within elements of a more complex data structure, such as a Knowledge Graph. Entity Alignment (EA) is the task of finding corresponding entities in two knowledge bases [4]. What differentiates our Record Matching problem from typical Record Linkage and Entity Resolution problems is the cross-domain correspondence between entities: records of Sound Recordings (SRs) and Musical Works (WKs). We cannot merge the datasets of

SRs and WKs as they have millions of entries and are conceptually distinct. Moreover the queried entity, in our case the SR, is provided with incomplete information, thus a preliminary EA has to be performed to retrieve the supplementary information of the SR (we call this process the enrichment of the SR). Because of this domain heterogeneity, our problem could be considered a multi-type EA as in [14], and could be compared to other problems in the literature as Metadata Cover Song Matching [3, 12], Medical Records Matching [5, 10], Academic pre-print paper marching [8] and person identity identification across various Social Media [1, 9].

Table 1 shows typical formats of SR and WK records. An SR record consists of metadata specifying attributes and their corresponding values: such as a title, a list of artists, a source provider and a register code as the ISRC (International Standard Recording Code), among others. A WK record is similar but contains, in replacement of the artist list, a list of contributors that are not directly linked with the artists and can include aliases, and a different register code as the ISWC (International Standard Musical Work Code). Titles can be expressed in various naming conventions, which might include characteristics of the SR (as for example being an instrumental recording), mentions of featured artists or composers, and can be in different languages, using different formats (abbreviations, versioning information). These variability renders traditional text similarity matching inadequate.

The artists and contributors in each record are listed with multiple names and aliases, necessitating the capability of list comparison. In addition, the attribute values may also influence the representation process, for instance, some source providers of the recordings do not provide a complete list of artists. The existing workflows in industry to solve this problem have demonstrated that the context of the record is crucial to make accurate predictions.

**Table 1.** Sound Recording (1~3) and Musical Work (4~5) examples

| title | artists | source | ISRC | |
|---|---|---|---|---|
| "Hold on to Me (feat. GTA)" | ["Jane","GTA"] | "spotify" | DEU241701542 | (1) |
| "Woah" | ["Wrongtom, Ragga Twins"] | "netease" | null | (2) |
| "Woah!" | ["The Ragga Twins", "Wrongtom"] | "tidal" | QZES82145030 | (3) |

| title | contributors | source | ISWC | |
|---|---|---|---|---|
| "HOLD ON TO ME" | ["托特", "claw yellow","Jane"] | "itunes" | T-123.456.789-1 | (4) |
| "Woah" | ["WRONGTOM MEETS RAGGA TWINS"] | "tencent" | T-071.443.120-5 | (5) |

Records (1) and (4) are a correct pair, despite 'GTA' appearing as a featured artist in the title and as an additional name in the artists list. Similarly, records (3) and (5) also match. However, the pair of records (2) and (5) differs; although the title of (2) is identical to (5) and its artists list aligns more closely with (5), it is non-matching as record (3) is registered with an ISRC code. We compile here a list of relevant challenges associated with the matching problem in this case:

a) Advance string similarity: titles and person names are provided with different aliases and naming conventions and can include extra information.
b) Fuzzy List comparison: artists and contributors names and aliases can partially match.

c) Attribute rules: source providers may not be reliable in providing complete information.
d) Contextual rules: an SR and WK may match depending on attribute values of related existing WKs.

We will see in Section 5 how our preliminary model addresses challenge (a) using the semantic similarity introduced by embeddings from pre-trained LLMs. Challenges (b) and (c) are dealt thanks to the list comparison rules that can be learned by a Transformer based model. Challenge (d) can be addressed by refining representations of records with their related ones, the so-called Context phase in Section 4.

### 2.1. The industry dataset

BMAT provided two datasets consisting of pairs of Sound Recordings (SR-SR) and Sound Recording to Musical Work (SR-WK). For instance, the SR to SR dataset contains thousands of annotated pairs (SR, SR, label), with the human annotated label indicating if they match, and similarly for SR to WK. The composition of the dataset is detailed in Table 2. In this paper, our focus lies on the SR to WK matching, which represents the most difficult problem and could be in principle solved purely through text metadata matching. The SR to SR matching, may require, as a final deciding step, to listen to the audio recordings of the corresponding records. In the case of SR to WK, human annotators may be also using complex rules or may have made their decision after querying other missing data in the internet, so the dataset may contain partial information to solve the classification problem at a 100% accuracy. This is because the information provided by different clients may contain incomplete attribute fields. To address this, BMAT has enriched all datasets by completing extra fields using software that relies heavily on historical matching data. In Table 3 an SR is enriched with contributors and their aliases, a WK record is enriched with artists and alt-titles.

**Table 2.** Industry Datasets: number of samples per benchmark and annotation.

| type | correct | related | different | unclear |
|---|---|---|---|---|
| **SR-SR** | 22116 (52.87%) | 16308 (38.99%) | 2866 (06.85%) | 537 (01.28%) |
| **SR-WK** | 29699 (66.36%) | 1061 (02.37%) | 13994 (31.27%) | - |

Industry datasets cannot be distributed as they contain sensitive client information, together with the enrichment that BMAT provides. This is the main reason to build a synthetic benchmark (see Section 3), which also aims to ensure the correctness of the labeling according to different rules of the musical domain.

**Table 3.** Sound Recording (7) and Musical Work (8) examples of the industry dataset

| title | artists | source | ISRC | *+ contributors* | |
|---|---|---|---|---|---|
| "Hold on to Me (feat. GTA)" | ["Jane","GTA"] | "spotify" | DEU241701542 | *+ ["Yellow Claw", "GTA"]* | (7) |

| title | contributors | source | ISWC | *+ alt-titles* | *+ artists* | |
|---|---|---|---|---|---|---|
| "HOLD ON TO ME" | ["托特", "claw yellow"] | "itunes" | T-123.456.789-1 | *+ ["Love","el amor"]* | *+ ["Jane Campbell]* | (8) |

## 3. A new benchmark

We introduce a synthetic dataset that tries to reproduce the challenges of the industry case but being generated from publicly available sources like `https://musicbrainz.org/`. From all the original attributes of the real industry case we have kept only the fundamental ones, without compromising the complexity of the problem and simplified others. For instance, the ISRC and ISWC codes of an SR and WK resp. are replaced with a boolean value denoting whether they are registered or not (see examples in Table 4).

**Table 4.** Sound Recording (9) and Musical Work (10) examples of the synthetic dataset

| **title** | **artists** | **source** | **- ISRC** | **+ *registered*** | |
|---|---|---|---|---|---|
| "Hold on to Me (feat. GTA)" | ["Jane","GTA"] | "spotify" | - DEU241701542 | + *True* | (9) |

| **title** | **contributors** | **source** | **- ISWC** | **+ *registered*** | |
|---|---|---|---|---|---|
| "HOLD ON TO ME" | ["托特", "claw yellow"] | "itunes" | - T-123.456.789-1 | + *True* | (10) |

We generated pairs of SRs and WKs following various rules derived from the industry case. We divide the generation rules into two groups: pairs independent of the context (we call them general Rules A and B) and pairs that are context-dependent (Rules 01 to 05). By 'context-dependent', we mean that to accurately validate the matching of the pair, the context of either the SR or the WK must be considered.

- Rule A: The titles and credits similarity are relatively high if the sample is correct; On the contrary, the pairs would have completely different names and low title similarity if the sample is different.
- Rule B: The titles are the same, and the credits similarity is low if the sample is correct. If the sample is incorrect, both the title and credits similarity are low.
- Rule 01 and 03: Generally, the SRs under these rules do not provide a satisfactory overall similarity, and the context offers a better match. Rule 01 focuses on credits similarity, while Rule 03 addresses both title and credits similarity.
- Rule 02: The context under this rule has very reliable overall similarity and is used as a reference to evaluate the target WK.
- Rule 04: The attribute value is decisive in this case. Both context and target WK provide average similarity, but context is registered. According to industry rules, the registered WK is considered the correct match.
- Rule 05: This rule also considers the attribute value as a dominant factor. The source of the record indicates that the record may have an incomplete artist list. Thus, the context record with a reliable source is used as reference for evaluation.

To construct the synthetic dataset, we generate matching and non-matching pairs for each rule. Both are required for training the neural network model and specifically for the Contrastive learning loss function that we use (see next section). For each rule we generate a proportion of examples that is indicated in Table 8. This is to mimic the characteristics of the real industrial dataset. The attribute values for each generated pair (such as title, artist, and contributor names) are sourced from real examples in the MusicBrainz dataset, which provides metadata of both SRs and WKs.

## 4. A preliminary model

We introduce here an overview of the method that we developed to solve the problems. It is intended to serve as a reference of the classification accuracy that can be achieved.

We provide a high-level description rather than details for reproducibility, as our focus is on the problem description and the industry and synthetic benchmarks.

We use a model based on a Transformer Neural Network to solve the classification problem which receives as input a Sound Recording (SR) and a Musical Work (WK), outputs a representation (embedding) of both, and the matching decision is made according to the distance of those representations (cosine similarity). The network is trained in two phases: Attribute phase and Context phase training.

In the Attribute phase we train an embedding model of both SRs and WKs that depends on the values of the attributes. An overview of the training process of the Attribute phase is shown in Figure 1. The model is given as input the encoded attribute values using a pre-trained text embedding model (analogous to a LLM), thus taking advantage of the semantic encoding capabilities of these pre-trained models. The model has two inputs: the SR and the WK (shown with two examples in the figure having the same title "Hello"). Both are input separately to two Neural Network blocks (represented here by the Embedding Block). The model representations of both the SR and WK, are trained following a Contrastive Learning approach [6] according to ground truth labels. If a matching input pair is provided then the representations are brought closer. If, on the contrary, a non-matching pair is provided, representations are brought further away. In the industrial case, we use a tool developed by BMAT called the Enricher to complete possible missing information of the entities before encoding. In Tables 3 and 4 we show examples of entity enrichment (adding alternative contributors, titles and artists). Each Embedding Block then encodes each attribute of the entities, along with enriched metadata, using a pre-trained LLM (green box in the figure). Attribute values are embedded together with their names, i.e., "Title: Hello" is the text sequence that is given as input to the pre-trained LLM and the encoded embedding $E_{title}$ is given as output. $E_{title}$, $E_{artist}$, $E_{source}$, $E_{contrib}$, ... are embedding vectors that can have different number of dimensions depending on the pre-trained model (768 in the case of a open-source BERT, or 1536 in the case of OpenAI text-embedding-ada-002). These embeddings are then fed into the Transformer Neural Network block (in grey in the figure), which consists of 6 Attention Layers, the typical building blocks of Transformers. Each blocks outputs the final encoding for the SR, $E_{SR}$, and the WK, $E_{WK}$. During this training phase, we further refine the SR and WK embeddings. The SR and WK are embedded separately, and a contrastive loss is applied to align the embeddings.

For the synthetic benchmark (introduced in Section 3), we don't use the Enricher in the Attribute phase training. Instead, in the Context phase, we employ other related SRs and WKs as contextual information in an attempt to complete the data. This approach serves as a substitute for the Enricher because it integrates more effectively with the embedding model. To establish a reference, we also evaluate a similar setup that excludes the use of contextual information.

After this training phase, the obtained encodings for the SR and WK, $E_{SR}$ and $E_{WK}$, are further refined by taking into account the context, specifically the related entity neighbours of the WK. While we don't provide technical details of this phase, we present some results in the next section.

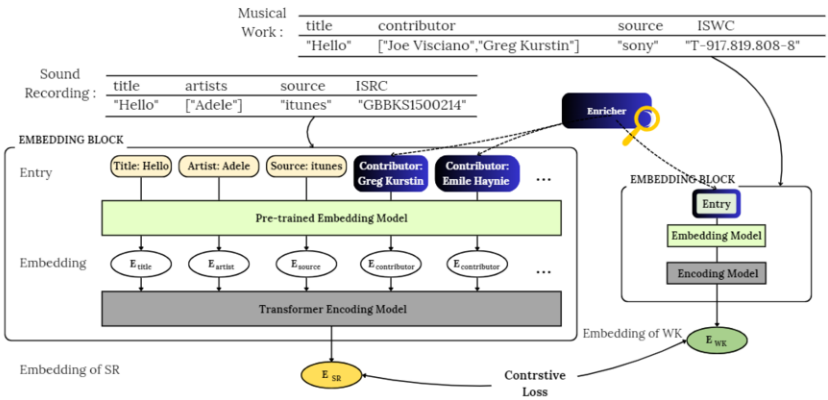


**Figure 1.** Training methodology and architecture of the model. An example Sound Recording (SR) and Work (WK) are provided to the model. Attribute types and values are encoded with a pre-trained Embedding Model and then input to the Transformer Encodeing Model which provides a final representation for both the SR, $E_{SR}$, and WK, $E_{WK}$. The model is trained with a Contrastive Loss from matching and non-matching pairs.

## 5. Preliminary results

We present the validation and benchmark results derived from both the industry (Section 5.1) and synthetic (Section 5.2) datasets, providing an evaluation and comparison across the usage of various pre-trained text embedding models (MiniLM, MPNet, text-embedding-ada-002). The Baseline model in this section is a minimal version model from BMAT. Section 5.1.1 provides specific examples from the industry dataset, showcasing challenges in aligning SR and WK metadata. In the synthetic benchmark results (Section 5.2), we compare the Attribute and Context phase training. The evaluation reveals the potential of rule extraction in noisy data environments.

**Table 5.** Evaluation of SR-SR validation on industry dataset

| Model | Label | precision | recall | f1-score | acc. |
|---|---|---|---|---|---|
| MiniLM | √ | 0.95 | 0.93 | 0.94 | 0.93 |
| | × | 0.92 | 0.93 | 0.93 | |
| MPNet | √ | 0.94 | 0.94 | 0.94 | 0.93 |
| | × | 0.92 | 0.93 | 0.93 | |
| text-embed -ada-002 | √ | 0.95 | 0.93 | 0.94 | 0.94 |
| | × | 0.92 | 0.94 | 0.93 | |
| Baseline | √ | 0.95 | 0.93 | 0.94 | 0.93 |
| | × | 0.92 | 0.94 | 0.93 | |

**Table 6.** Evaluation of SR-WK retrieval on industry dataset

| Model \ Hits@ | 1 | 5 | 10 | 20 | 50 | 100 |
|---|---|---|---|---|---|---|
| BM25 | 59.6 | 75.7 | 80.4 | 84.0 | 87.4 | 89.0 |
| MiniLM+Our | 75.9 | 88.1 | 89.9 | 91.2 | 92.8 | 94.1 |
| MPNet+Our | 80.9 | 92.2 | 93.3 | 94.2 | 95.3 | 96.1 |
| ada-002+Our | 86.3 | 96.7 | 97.4 | 97.8 | 98.4 | 99.1 |
| Baseline | - | 87.4 | 91.5 | 94.0 | 95.6 | 96.1 |

**Table 7.** Evaluation of SR-WK validation on industry dataset

| Model | Label | precision | recall | f1-score | acc. |
|---|---|---|---|---|---|
| MiniLM | √ | 0.82 | 0.78 | 0.80 | 0.75 |
| | × | 0.62 | 0.67 | 0.64 | |
| MPNet | √ | 0.76 | 0.80 | 0.78 | 0.70 |
| | × | 0.57 | 0.51 | 0.54 | |
| MiniLM Finetuned | √ | 0.91 | 0.87 | 0.89 | 0.86 |
| | × | 0.77 | 0.84 | 0.80 | |
| MPNet Finetuned | √ | 0.87 | 0.90 | 0.88 | 0.85 |
| | × | 0.79 | 0.74 | 0.77 | |
| MiniLM+Our | √ | 0.88 | 0.91 | 0.89 | 0.85 |
| | × | 0.81 | 0.75 | 0.78 | |
| MPNet+Our | √ | 0.88 | 0.93 | 0.91 | 0.87 |
| | × | 0.85 | 0.76 | 0.80 | |
| text-embed -ada-002+Our | √ | 0.89 | 0.94 | 0.92 | 0.89 |
| | × | 0.87 | 0.79 | 0.83 | |
| Baseline | √ | 0.84 | 0.96 | 0.89 | 0.85 |
| | × | 0.89 | 0.63 | 0.74 | |

### *5.1. Validation of the Industry dataset*

The best results of our approach applied to the industry dataset came from using 3 to 6 Attention layers with 4 to 8 attention heads in the Transformer Block (grey block of Figure 1). The specific setup of the model depends on the features' dimensions.

Compared to BMAT baseline, we have improved the performance without field-wise comparison (e.g. SR title to WK title comparison), and process all attribute values in an unordered sequence, which enables this approach to be generalized to other knowledge-based tasks.

The Hits@5 and Hits@10 scores of our model significantly outperform the baseline. Although the Hits@1 score is not available for the baseline, it should reflect the accuracy of the embedding search within the database. The Hits@20, Hits@50, and Hits@100 scores are on par with BMAT baseline. In summary, the structure of the candidate list has been optimized, resulting in a higher ranking of related knowledge. However, the capability to retrieve less similar knowledge has only seen a marginal improvement.

#### *5.1.1. Examples from the industry dataset*

The following cases were successfully identified by our method but were missed by BMAT baseline model.

In the case of different languages, we observed instances such as the artist "Magic Bell" being represented as "Clopotelul Magic," where "Clopotelul" means bell in Romanian. Similarly, the Russian title " Без тебя я не я " (I am not me without you) was romanized as "BEZ TEBJA JA NE JA," and the Korean title " 다시 너를 " (I'll put you back on) was translated to "Once again."

Background knowledge played a crucial role in aligning artists and contributors, as seen with "Kenshi Yonezu" and its Japanese original name " 米津玄師 " or the Chinese contributors " 詹于萱 " aligning with her group "BY2". Another example includes " 許冠傑" and "張國榮" (Leslie Cheung) being aligned with "Sam Hui" and contributors "Kai Sang Chow | Leslie Cheung | Sam Hui."

Semantically similar cases were also identified, such as the title "The Buzz (The Brave Brass Remix) [Instrumental]" aligning with "The Buzz (The Brave Brass Remix) [No Drums & No Vocals]," where "No vocals" implies "Instrumental."

Miscellaneous factors further complicated the alignment process. For instance, abbreviations in titles like "Get Ready (Inst.)" corresponded to "GET READY (INSTRUMENTAL)." Phonetic transcriptions of Chinese titles, such as "Yi Qi Zhou Guo De Ri Zi (Album Version)," matched with "一起走過的日子- 電影"至尊無上II之永霸天下" 歌曲." Different formatting styles were evident in titles like "U.F.O. (K-Hole Riddim)" and "UFO." Additionally, unknown phonetic transcriptions such as "Chin Tien" were aligned with the Chinese title "晴天 "

These examples demonstrate the complexity of aligning metadata across various languages, the importance of background knowledge, semantic similarities, and various miscellaneous factors, underscoring the challenges inherent in ensuring accurate and consistent metadata alignment in SRs and WKs.

### *5.2. Synthetic benchmark results*

For our tests, which include model training and evaluation using the synthetic benchmark, we utilized text embeddings from MPNet (an open source pre-trained model).

**Table 8.** Synthetic Dataset (SR-WK): number of samples per rule and annotation.

| rule | correct | different | total (percentage) |
|---|---|---|---|
| **A** | 6827 | 7028 | 13855 (36.87%) |
| **B** | 6502 | 7393 | 13895 (36.98%) |
| **01** | 959 | 946 | 1905 (5.07%) |
| **02** | 985 | 954 | 1939 (5.16%) |
| **03** | 934 | 966 | 1900 (5.05%) |
| **04** | 999 | 1047 | 2046 (5.44%) |
| **05** | 1020 | 1020 | 2040 (5.43%) |
| **total** | 19354 (51.50%) | 18226 (48.50%) | - |

**Table 9.** Attribute and Attribute + Context training performance

| Rule | Attribute phase | Attribute+Context phase |
|---|---|---|
| A, B | 82.91 | 82.88 |
| A, B, 01 | 78.21 | 80.45 |
| A, B, 02 | 79.89 | 81.67 |
| A, B, 03 | 81.78 | 82.30 |
| A, B, 04 | 79.16 | 80.91 |
| A, B, 05 | 79.62 | 81.55 |
| A, B, (01...05) | 75.99 | 80.80 |

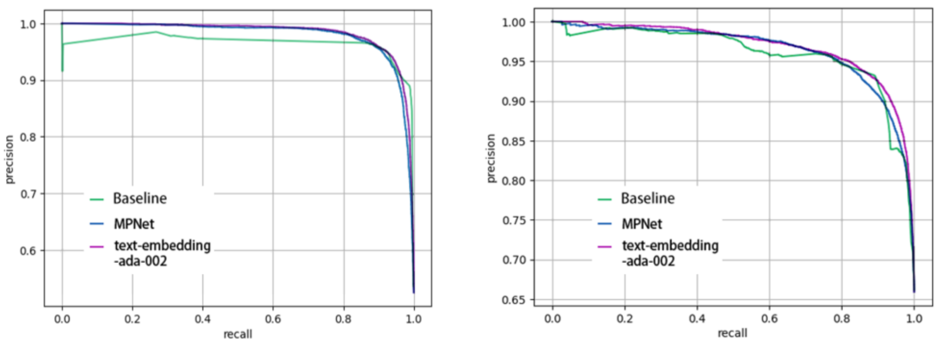


**Figure 2.** Precision-Recall (PR) curves. Classification results of the industry benchmarks: SR-SR (left) and SR-WK (right). In green we plot the PR curve of the BMAT baseline. The other two PR curves correspond to the model presented here using two different pre-trained embedding models for encoding text attributes: MPNet (blue) and the OpenAI text-embedding-ada-002 (red). We can see that our model performs better in general having a greater area under the curve. Although it can hardly be seen in the plot, the OpenAI text embeddings perform better.

The significant drop in performance (see Table 9) when including contextual rules 01 to 05 (the biggest drop being from 82 to 75 accuracy) indicates the inefficiency of the current method when dealing with context. But the improvements when including the Contextual training phase (shown in the second column of Table 9) suggest the potential for rule extraction despite the presence of noisy data.

## 6. Conclusions

We have presented a new record linkage industrial problem in which Sound Recordings (SRs) text metadata have to be linked with their corresponding Works (WKs). The problem is challenging and requires complex string similarity of titles and names, together with the correct application of complex rules of the music industry domain.

As the real industry problem is complex and noisy and contains proprietary data we have constructed a synthetic dataset trying to include all relevant challenges present in the original industry problem. We focus the contribution of the paper on this benchmark that will be made available soon.

Finally we have presented preliminary results to both the real industry case and the synthetic benchmark, introducing a model that leverages from pre-trained embedding models encoding titles and names, and being able to compare them at a semantic level. The model is a transformer neural network architecture that is trained in two phases with a contrastive loss and is able to take into account the context of each record (SR and WK). Although we see an improvement in the classification when the relational (contextual) information is taken into account, we are not close from achieving the performance that a human annotator would achieve.

## Acknowledgments

Research supported by Red.es (Ministerio para la Transformación Digital y la Función Pública) within the Recovery Plan, transformation and Resilience through the MatchingLearning project, reference C005/21-ED2021/C005/00149157, co-funded by the European Commission – NextGenerationEU.

## References

[1] William M Campbell, Lin Li, C Dagli, Joel Acevedo-Aviles, K Geyer, Joseph P Campbell, and C Priebe. Cross-domain entity resolution in social media. *arXiv preprint arXiv:1608.01386*, 2016.

[2] V. Christophides, V. Efthymiou, G. Palpanas, T.and Papadakis, and K. Stefanidis. An overview of end-to-end entity resolution for big data. *ACM Computing Surveys (CSUR)*, 53(6):1–42, 2020.

[3] Albin Andrew Correya, Romain Hennequin, and Mickaël Arcos. Large-scale cover song detection in digital music libraries using metadata, lyrics and audio features. *arXiv preprint arXiv:1808.10351*, 2018.

[4] Nikolaos Fanourakis, Vasilis Efthymiou, Dimitris Kotzinos, and Vassilis Christophides. Knowledge graph embedding methods for entity alignment: experimental review. 37(5):2070–2137, 2023.

[5] Erel Joffe, Michael J Byrne, Phillip Reeder, Jorge R Herskovic, Craig W Johnson, Allison B McCoy, Dean F Sittig, and Elmer V Bernstam. A benchmark comparison of deterministic and probabilistic methods for defining manual review datasets in duplicate records reconciliation. 21(1):97–104, 2014.

[6] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021.

[7] Manuel Leone, Stefano Huber, Akhil Arora, Alberto García-Durán, and Robert West. A critical re-evaluation of neural methods for entity alignment. 15(8):1712–1725, 2022.

[8] J. Lin, Y. Yu, Y. Zhou, Z. Zhou, and X. Shi. How many preprints have actually been printed and why: a case study of computer science preprints on arxiv. *Scientometrics*, 124(1):555–574, 2020.

[9] Danna Lu, Qing Li, Zikang Liu, Shumei Li, Xiaobo Wu, and Shudong Li. A review of user identity linkage across social networks. In *2023 8th International Conference on Data Science in Cyberspace (DSC)*, pages 429–436. IEEE, 2023.

[10] Mahin Ramezani, Guru Ilangovan, and Hye-Chung Kum. Evaluation of machine learning algorithms in a human-computer hybrid record linkage system. 2846(4), 2021.

[11] Yuting Song, Biligsaikhan Batjargal, and Akira Maeda. Cross-language record linkage based on semantic matching of metadata. 17(1), 2019.

[12] F. Yesiler, Ch. Tralie, Albin A. Correya, Diego F. Silva, Ph. Tovstogan, E. Gómez Gutiérrez, and X. Serra. Da-TACOS: A dataset for cover song identification and understanding. 2019. Accepted: 2019-11-07T09:07:22Z Publisher: International Society for Music Information Retrieval (ISMIR).

[13] Kaisheng Zeng, Chengjiang Li, Lei Hou, Juanzi Li, and Ling Feng. A comprehensive survey of entity alignment for knowledge graphs. 2:1–13, 2021.

[14] Qi Zhu, Hao Wei, Bunyamin Sisman, Da Zheng, Christos Faloutsos, Xin Luna Dong, and Jiawei Han. Collective multi-type entity alignment between knowledge graphs. In *Proceedings of The Web Conference 2020*, pages 2241–2252. ACM, 2020.