CEUR-WS.org/Vol-2949/paper5.pdf

Masked Language Model Entity Matching for Cultural Heritage Data

Dominique Piché[1], Amal Zouaq[1], Michel Gagnon[1], and Ludovic Font[1]

École Polytechnique de Montréal
2500 chemin de Polytechnique, Montréal, QC, Canada
{dominique.piche, amal.zouaq, michel.gagnon, ludovic.font}@polymtl.ca

Abstract. Entity resolution is a well-known issue in Cultural Heritage data integration, as existing metadata for cultural works is typically distributed across multiple databases maintained by various actors. Identifying classes of equivalent entities is thus a non-trivial but necessary task. Entity linking between heterogeneous data sources is made complex by variations in data schemes and cleanliness. In this work, we test masked language models (MLM) fine-tuned for entity matching on real-world literature data. We examine the impact of pre-processing and input annotation strategies. Our results show that MLMs outperform or match our rule-based heuristics in most scenarios. Interestingly, the impact of the chosen MLMs language, data pre-processing and input annotation all have little to no effect on matching performance in our experiments.

Keywords: Entity resolution · Knowledge base · Transformers · LRM · Cultural heritage.

1 Introduction

The accumulation of information through dispersed data stores in the past decades has motivated the development of interlinked knowledge graphs able not only to represent complex information through graph-based structures, but also to link information hosted on servers operated by unrelated entities. The translation of various data stores describing a given domain into one unified knowledge graph requires the detection of duplicate representations of the same real world entities, as each real-world entity having a single unique identifier allows for the aggregation of data describing it, and references to be made to it from external sources. This problem is known as entity matching, alignment or resolution. Several issues complicate matching entities between data sets, notably the absence of preexisting global identifiers and variations in data cleanliness, data schemes and database scopes. Cultural heritage (CH) data is particularly prone to these types of problems due to the way it is collected and maintained, as discussed below. Our aim is to determine to which extent neural solutions mitigate the need for data cleaning and restructuring before match prediction.

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

1.1 Cultural heritage data integration

The integration of CH metadata into linked knowledge bases has been the focus of various initiatives, as various cultural domain actors such as national libraries, museums and publishers. Cultural Heritage (CH) data describes objects and works of cultural significance and their contents, as well as the context of their creation. Most existing data sets focus on physical items, e.g. museum data describing artefacts. These data sets are typically constituted through manual cataloging. As CH data focuses heavily on physical manifestations of works, matching identical conceptual works realized in separate manifestations is required. For example, two separate data sets may describe separate editions of the book Don Quixote in different languages, yet the thematic information of the original conceptual work should be aggregated in a single entity through entity matching.

A recurring problem for translation of collections of real-world CH data sets into unified knowledge bases is the absence of global unique identifiers for entities. The ideal method for matching entities between sources would be the use of identifiers such as an International Standard Book Number, or ISBN, for book editions. In practice, however, most entity types are only described with internal unique identifiers, requiring alternatives for cross-source matching. In the case of literary data, it is rare to find a global unique identifier for entities such as authors, even though ISNI and VIAF is slowly becoming more widespread. Existing data sources rarely make the distinction between conceptual works and their manifestations, and, consequently, works are left without global identifiers. Evolving cataloging standards and data entry practices often mean that not only do sources often use different naming and encoding conventions for information such as titles and dates, but that data sets contain internal inconsistencies.

In the absence of identifiers, the next best method for resolving entity duplication is the detection of similar entity representations, i.e. entities with enough attribute similarity for confident matching. Implementation of rule-based matching has its challenges, notably data cleanliness and heterogeneity in data schemes. Evaluating the entity similarity within or between sources is difficult, as important attributes such as titles and names are formatted according to different rules from record to record, even within one data set.

Extraction and cleaning of source record data before translation into a single model can do much to solve this "dirty entity" problem, allowing for comparison of pre-processed and clean representations of entitiesThis process is labour intensive, and requires extensive domain knowledge in order to map information between models, as multiple versions of a cataloging standard present within single source, duplicated attributes, typos and mislabeled values are frequent occurrences. Thus, finding solutions less reliant on pre-processing may speed up development time.

1.2 Neural solutions for matching dirty entities

As these challenges are a frequent occurrence amongst most CH data integration initiatives, the development of a generalizable solution for entity matching without the need for domain-specific extraction, cleaning and matching rules could help speed future knowledge base creation efforts, especially in cases with inconsistent source data. Tools leveraging neural methods for entity matching have been developed in the past years, with graph-embedding based neural models such as [2] currently some of the best performing on Knowledge Graphs.

On the other hand, pre-trained Masked Language Models (MLMs) taking texts as inputs, such as BERT, have outperformed existing state-of-the-art methods on various language understanding tasks, and their application to the entity matching problem has beaten records on benchmark data sets [1,9]. As CH data often contains large amounts of textual descriptions, the application of MLMs to the matching task could potentially accelerate development and increase performance.

We explore how MLM methods can be used to unify four literature metadata sets from various actors of Quebec's literary world into a homogeneous knowledge base. As our source data are mostly in French, pre-trained French language models are among those leveraged to generate entity pair embeddings for match prediction. In particular, we focus on Work and Author entities, central types in literary metadata. We compare MLMs to a string similarity baseline on fully cleaned and labeled data. We then test language models on various input formats in order to determine how much data pre-processing and annotation is required for peak performance. Our key research questions are:

1. How do masked language models compare to rule-based and domain-related heuristics in the context of CH data?
2. What is the impact of pre-processing and input formats on MLM matcher performance?

The paper is structured as follows. We present an overview of our source data and target ontology in section 3, our global information extraction, cleaning, matching and translation process in section 4.1, with particular focus on the entity matching phase. Our MLMs are described in section 4.2, and our training sets and baselines in sections 4.3 and 4.4 respectively. We present and analyse our results in sections 5.

2 Related Work

Alignment of instances between data sets is a well known research problem [1,3, 9,11,12,14]. Initial approaches were mostly rule-based but they require domain expertise and maintenance in order to be adapted to new data [3]. Probabilistic matching of instances then remained the primary method for alignment, notably in Semantic Web applications, as in [14].

Further development of neural deep learning approaches has seen them applied successfully for entity matching, with some top performing models using

Recurrent Neural Networks with Long Short Term Memory [5, 6]. Recent surveys [11] mention these types of architectures as achieving state-of-the-art performance on benchmark entity matching data sets such as DBLP-ACM and Walmart-Amazon [4].

More recently, Natural Language Processing (NLP) techniques using transfer learning through transformers pre-trained on masked language modeling tasks demonstrated state-of-the-art performance on language-related downstream tasks. As entity matching can be seen as the binary classification of two sequences as identical or not, these models can be easily fine-tuned on this task. Two existing works [1, 9] using this method have already outperformed existing models on the Magellan alignment benchmarks. The second, Ditto [9], makes use of data augmentation and attribute annotation strategies challenging the language model to learn nuances for embedding similar sequences. Both solutions add a fully connected layer and a SoftMax classification layer for match predictions on top of pre-trained English language models (BERT, DistilBERT, RoBERTa or XLNet).

In this work, given that our data sources are in French, we propose a matching architecture based on French pre-trained MLMs. Two MLMs built on RoBERTa's architecture are of interest. CamemBERT [10] is trained on the French portion of the OSCAR corpus, a pre-filtered version of Common Crawl, while FlauBERT [8] trains on a combination of French sets, including WMT19, OPUS, Wikimedia and Project Gutenberg.

3 Data overview

Our experiments are performed in the context of a project headed by the Ministry of Culture and Communications of Quebec (MCCQ), which aims to create a knowledge base for Quebec's literary data, sourced from the domain's stakeholders. Data stems from four sets, provided by the Quebec national library and archives (BAnQ), Messageries ADP (a book distributor), the Infocentre Littéraire des écrivains du Québec (ILE), and Les Éditions Hurtubise, a Quebec publisher.

3.1 Source records

The provided sets describe literary entities organized in records. Records consist of book metadata descriptions, with one entry generally describing a specific edition of a book, its content, physical description, publication and author.

Our sources present the typical challenges for entity resolution in CH data. Although the sets have similar scopes, there are minor differences regarding the content of a single record. A book series may be described in a single record, as is the case in BAnQ, or be distributed over multiple records, as is the case in ADP's data. Conventions for encoding titles and names also vary: some sources remove leading pronouns, some place last names before first names for authors, some split titles and subtitles into separate fields, etc. Records have varying degrees of quality; some only contain uppercase text, some are cut off after a certain number of characters. Entity attributes vary, with, for example, BAnQ being the only set containing authors' dates of birth. Frequently, attributes defined in a source's data scheme are absent a large part of records. Finally, some sources have separate collections of records for authors and books, while other sources concatenate author and book information into single records.

To train our MLM matchers, labeled training sets of positive and negative matches must be generated. One unique global identifier, the ISBN, is available across data sets and can be used to generate these sets. We present the generation and content of these sets further in this section.

3.2 Ontological Model

Our final model uses a linked data structure, defined in an ontology, with data being stored in RDF triples. As we are working with literary data, we chose to implement an IFLA Library Reference Model (LRM) [13] based knowledge graph as a target model.

LRM is an implementation-agnostic conceptual model for representing library data, and was conceived to replace a series of previous reference models, including the Functional Requirements for Bibliographic Records (FRBR) [7]. LRM defines 11 types of entities in a hierarchical structure, with some schema elements retained from the previous reference models.

Works, **Expressions** and **Manifestations** represent different layers of abstraction for intellectual works and their physical manifestations. The *Work* is the higher conceptual level, representing the work of art as created by the writer (the story of "Alice in Wonderland"); the *Expression* represents the embodiment of that work in a written form (the French translation by André Gagnon); the *Manifestation* represents the creation of a series of physical entities that correspond to published editions of that text (the 2012 edition published by Hurtubise). The *Item* is disregarded, as our model is not being developed for inventory management.

Nomen entities encompass identifiers, names, titles, terms, descriptors and subject headings. The existence of this type is justified by the necessity to represent assignation relationships between *Agents* and *Nomens* and identification of other characteristics of individual *Nomens* such as schemes and encoding language. In the context of fusing data sets using varied classification schemes, this functionality is crucial.

4 Methodology

We describe the data model in section 4.1 and the structure of our transformer-based matching model in section 4.2. The structure of and generation strategy for our labeled training data sets is laid out in subsection 4.3, and heuristic matching baseline, experiments and metrics in 4.4.

4.1 General Architecture for Knowledge Base Extraction

The creation of a unified knowledge base from disparate sources generally involves a set of data processing steps. These steps have an incidence on the performance of the subsequent modules (e.g. entity matching). Particularly, cleaning attributes and aligning schemas across data sets can help facilitate rule-based entity matching. However, the success of each of these steps is highly dependent on data characteristics as outlined in section 3.1. We developed a multi-stage pipeline for extracting, cleaning, enriching, aligning and translating entities from source records into our LRM-based model, with experiments focusing on the alignment phase.

Six main phases compose our pipeline. Records are extracted from sources and given unique identifiers in **Phase 1**. Entities contained in records are extracted in **Phase 2**. A record typically contains one Work entity, with associated Expression and Manifestation, one Author and one Publisher. The entities are restructured into a common intermediate representation based on our LRM model. For example, a publication description containing "354 p." is assigned to a Manifestation entity's page number attribute. In **Phase 3** that the content of attributes is cleaned and standardized. Errors or special filing characters are identified and removed. Sub-attributes, such as first names and surnames for names, are extracted. In **Phase 4**, entities are enriched through external resources; language strings are replaced with links to WikiData language pages, place names are replaced with place entities with unique codes and organized into hierarchies, etc. Classes of equivalent entities are identified in **Phase 5**. These equivalence classes' canonical representations, generated from merging entities that make up the classes, are translated into the target graph in **Phase 6**.

The matching phase in more detail. Conceptually, the task is to identify clusters of local entities representing the same real-world entity, with the entity matcher - the MLM - determining whether representations of two clusters imply these clusters should be merged. Once a cluster is newly modified, it is then compared to other clusters once more to check for new possible matches. This process allows for gradual integration of further data sets. Once the pool of clusters reaches a stable state (no link between clusters can be found), translation to the final graph can start.

4.2 Masked Language Model for Entity Matching

Our architecture is inspired by recent works [9] and [1], but differs by the nature of data (database records, semi-structured data), the language (French) and our experiments on the impact of input formats and pre-processing strategies. The MLM is enriched with a fully connected layer and a SoftMax classification layer added in the output layers. Fine-tuning and evaluation is performed on the train, test and validation sets presented in section 4.3. The pre-trained models selected for generating sequence embeddings for entities are CamemBERT [10] and FlauBERT [8], both based on RoBERTa. Both models are pre-trained on a masked language modeling task on large French corpora. We use parameters from previous works [1] for batch size (32) and learning rate (3e-5).

Input format. BERT-like language models, such as the ones we use, take text as input, composed of one or two sequences, and, for a classification task, a label to be predicted. In our case, each entry represents a pair of entities to be aligned, composed of two entity strings and a label identifying whether they are to be aligned or not. Our labeled sets represent entries on one line, with a tab separating each of the three elements: the string for entities 1 and 2, and the label.

Table 1. Input formats for different pre-processing and annotation strategies

	Pre-processing and annotation	Structure
1	RegEx cleaning and data annotation with LRM attribute names, values and special tokens	[C] attribute1 [V] value1 [C] attribute2 [V] value2 [...] [C] attributeN [V] valueN
2	Data annotation with original schema names, values and special tokens, without cleaning	[C] attribute1 [V] value1 [C] attribute2 [V] value2 [...] [C] attributeN [V] valueN
3	Data annotation with original schema names and values, without cleaning or special tokens	attribute1 value1 attribute2 value2 [...] attributeN valueN
4	Raw data values without cleaning or annotation	value1 value2 [...] valueN

We experiment with four input formats differing in pre-processing and annotation strategies, in our training sets. These input formats concatenate extracted attribute values into strings as entity representations. Annotation strategies add special tags (e.g. [C]) that indicate the schema/meaning of a given text token to help the matcher identify the various elements in the string. The first format is based on what was proposed by Li et al, 2020 [9]. The data in this format is cleaned with regular expressions normalizing punctuation, removing special characters and structuring strings in the same way (e.g. firstname then lastname) among data sets. The data is restructured to follow our LRM schema. Input strings consist of alternating pairs of attribute names and values, separated by special tokens indicating whether the following substring is the name of an attribute ([C]) or its value ([V]). Attribute names are one or two characters long: t for title, st for subtitle, a for Author name, etc. The second format is similar, but with the special tokens removed. Attribute names and values are separated by a space only. The third format does away with cleaning and restructuring. Entries are created from original attribute names and values, only separated by a space. Attribute names can thus be a textual label or a standardized field identifier, such as a MARC21 code, depending on the source. The final input format omits attribute names and cleaning entirely, and is a concatenation of original attribute values only. Examples of complete entries for these formats are shown in Table 2.

Table 2. Examples for input formats of a positive match, with sequence 1 from BAnQ and sequence 2 from Hurtubise ([C]: Column identifier token, [V]:Value identifier token, t: title, a: author, 245a: MARC21 subfield for title, etc.)

#		Value
1	Seq 1 :	**[C] t [V]** Être un héros **[C] e [V]** La Courte échelle **[C] st [V]** des histoires de gars **[C] lp [V]** Montréal **[C] np [V]** 218
	Seq 2 :	**[C] t [V]** Être un héros **[C] ap [V]** 2011 **[C] a [V]** Simon Boulerice **[C] e [V]** La Courte échelle **[C] st [V]** des histoires de gars **[C] lp [V]** Montréal **[C] np [V]** 218
2	Seq 1 :	**[C] 245a [V]** Être un héros : **[C] 245b [V]** des histoires de gars / **[C] 260b [V]** La Courte échelle, **[C] 300a [V]** 1 ressource en ligne (218 p.) : **[C] 260a [V]** Montréal :
	Seq 2 :	**[C] 0 [V]** Être un héros : des histoires de gars **[C] 2 [V]** Boulerice, Simon **[C] 3 [V]** La Courte échelle, 2011, 218 p. **[C] 1 [V]** 2011 **[C] 4 [V]** Montréal
3	Seq 1 :	**245a** Être un héros : **245b** des histoires de gars / **260b** La Courte échelle, **300a** 1 ressource en ligne (218 p.) : **260a** Montréal :
	Seq 2 :	**0** Être un héros : des histoires de gars **2** Boulerice, Simon **3** La Courte échelle, 2011, 218 p. **1** 2011 **4** Montréal
4	Seq 1 :	Être un héros : des histoires de gars / La Courte échelle, 1 ressource en ligne (218 p.) : Montréal :
	Seq 2 :	Être un héros : des histoires de gars Boulerice, Simon La Courte échelle, 2011, 218 p. 2011 Montréal

4.3 Training and evaluation sets

We need evaluation sets for the entities we seek to align, in our case Works and Authors. Each set used, shown in Table 3, is comprised of pairs of entities with labels 1 for positive pairs and 0 for negative pairs.

Positive pairs for Works are identified using the only unique ID available across sources: the ISBN of the Manifestations associated with the Works. The cardinality of the relations between Works and Manifestations allots the inference that Works having Manifestations with the identical ISBNs are refer to the same entity. As for Authors, without a unique identifier (only one source contains ISNI or VIAF IDs), we must assume that two Author records that have written the same Work and have similar names (using Levenshtein ratio 1) are the same.

Table 3. Descriptive statistics of our training, validation and test sets

Entity	Train (80%)	Valid (10%)	Test (10%)	Positives	Negatives
Work	44 573	5 572	5 574	18 573	37 146
Author	34 645	4 331	4 332	14 436	28 872

$$Levenshtein(str1,\ str2) = \left(1 - \frac{string\ edit\ distance}{max(len(str1),\ len(str2))}\right) * 100 \qquad (1)$$

For negative pairs, pairs of records are chosen at random. If two selected records do not have a shared ISBN and there is a large edit distance between their names or titles or they have similar titles yet fundamentally incompatible characteristics such as different volume numbers, then they are not considered to be identical entities in the real world. To avoid too great an imbalance between the amount of positive and negative pairs, we limit the amount of negatives to twice the number of positives. Negative and positive pairs are acquired and separated into training, validation and tests sets containing identical proportions of positive and negative pairs.

As ISBNs were used as a key for the creation of the training, validation and test sets, they cannot be part of entity encodings, as since the task at hand is finding pairs of entities that do not have shared unique identifiers, this would unduly influence our results.

4.4 Baselines and evaluation metrics

Our baseline alignment method is based on string similarity between attributes describing the entities. Using only entity names may be problematic, as some non-matches may share more similar names than some true matches (as shown in Fig. 1, which plots Levenshtein ratios. 1 between Work titles). The left figure shows a global view of ratios, while the right one illustrates the overlap between similarities of positive and negative pairs in the 60-80 range that make determining a clean threshold impossible.

Relying only on titles will cause false positives; comparing the names of the authors of each candidate entity eliminates these cases. If two works have very similar author names in addition to having similar titles, then the rule-based method considers them to be a match. Similarly, authors are matched if they are similarly named (L1 > 95) and wrote similar books (L1 > 90). Standard recall, precision, accuracy and F1 score metrics are employed. Results are presented in section 5, examined in section 6.

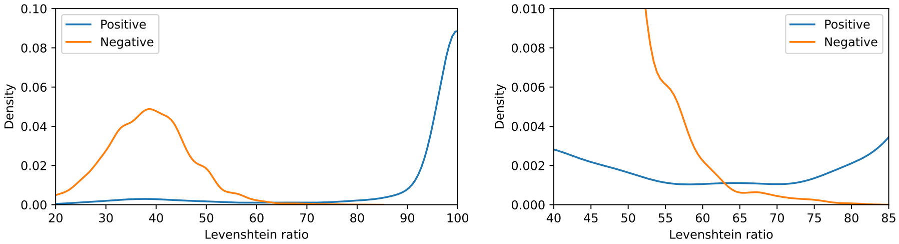

Fig. 1. Density estimate of Levenshtein ratios of titles of identical (positive) and distinct (negative) Works using format 1 (Table 1)

5 Results

Table 4 shows our experimental results for each of our research questions. In section A, we compare the performance of our best matching model (determined in later experiments B, C and D) with our heuristic-based baseline on completely cleaned, LRM-structured data, annotated with special tokens (see example in Table 2). These results show MLMs combined with SoftMax classifiers can vastly outperform domain-aware rules, while illustrating that Author matching is trivial in comparison with the Work matching task; a more challenging test set is required for better comparisons. However, our best heuristics are unable to achieve this perfect result, with MLMs having better performance on edge cases.

Training separate models for different entity types, however, uses more computational resources. We test whether a single model trained on joint Author-Work sets can reach similar levels of performance. Section B of demonstrates that a single model performs just as split models. Section C compares the performance of both pre-trained French MLMs. CamemBERT outperforming FlauBERT on every metric except precision, we use this model on all input formats 2 in section D. Given the equal or higher F1 scores (Table 4's section D) for drastically reduced pre-processing, these costly steps may be omitted without significant risk. More pre-processing steps require developing domain-specific rules with poor reusability and require active maintenance in the case of addition of data. Our results show that MLMs allow us to avoid this problem for entity matching in our context, as shown by format 4's results.

6 Conclusion and Further Research

We propose a MLM entity matching model for matching digital cultural records expressed in French, confirming that transformer-based pre-trained masked language models are a powerful tool for entity matching for cultural heritage data. We conclude that high performances can be achieved through fine-tuning even on very limited, unprocessed and heterogeneous labeled data. Even if our heuristic methods obtain very high scores, their comparable labour cost and poor

Table 4. Aggregated results

A. **Best matcher** versus **baseline**			**Evaluation on test set**			
Model Type	Peak Ep.	Entity	F1	Recall	Precision	Accuracy
Rule-based		Author	0.9955	0.9965	0.9945	0.9970
		Work	0.9119	0.8407	0.9962	0.9458
CamemBERT	4	Author	**0.9986**	**0.9979**	**0.9993**	**0.9991**
	4	Work	**0.9978**	**0.9968**	**0.9989**	**0.9986**
B. **Joint Author-Work matcher**						
Model name	Peak Ep.	Max Ep.				
Rule-based			0.9501	0.9088	0.9954	0.9682
CamemBERT	10	10	**0.9991**	**0.9988**	**0.9994**	**0.9994**
C. **MLM arch.** on Author-Work set						
Model name	Peak Ep.	Max Ep.				
CamemBERT	10	10	**0.9991**	0.9988	**0.9994**	**0.9994**
FlauBERT	10	10	0.9989	**0.9991**	0.9988	0.9993
D. **Pre-processing formats**						
Input format	Peak Ep.	Max Ep.				
1	10	10	0.9991	0.9988	0.9994	0.9994
2	4	10	0.9992	0.9988	**0.9997**	0.9995
3	4	10	0.9993	0.9991	0.9994	0.9995
4	10	10	**0.9994**	**0.9994**	0.9994	**0.9996**

reuse make MLM matching models a competitive alternative, leaning matching rules automatically. The use of heterogeneous data models has little impact on MLM performance in our tests, facilitating integration of new data sets. Furthermore, our results show that pre-processing does not improve MLM matching over unprocessed data, meaning development cost may be significantly reduced. Performance differences between unprocessed data sets and those annotated as suggested in Li et al. [9] were minimal; tagging of values with column names and special tokens may not yield performance substantial improvements.

One limitation is the use of ISBNs for positive pair generation. Correlated ISBNs mean correlated publication data; this may hamper the model's ability to correctly match works published in different places, times or languages. As positive pair generation for authors also relies on shared publications, they may suffer the same issue. Another possible bias is introduced by negative pair generation, as in order to generate high confidence negative matches, very strict rules are employed, precluding the presence edge-cases training sets.

In future work, we will further investigate the results of the English MLM for our task and we will try to replicate our results on data pre-processing and annotation on other entity matching data sets, rework negative pair annotation for more challenging examples, as well as test competing graph-based entity clustering methods.

References

1. Brunner, U., Stockinger, K.: Entity matching with transformer architectures-a step forward in data integration. In: International Conference on Extending Database Technology, Copenhagen, 30 March-2 April 2020. OpenProceedings (2020)
2. Chen, M., Tian, Y., Yang, M., Zaniolo, C.: Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. arXiv preprint arXiv:1611.03954 (2016)
3. Churches, T., Christen, P., Lim, K., Zhu, J.X.: Preparation of name and address data for record linkage using hidden markov models. BMC Medical Informatics and Decision Making **2**(1), 1–16 (2002)
4. Das, S., Doan, A., Psgc, C.G., Konda, P., Govind, Y., Paulsen, D.: The magellan data repository (2015)
5. Ebraheem, M., Thirumuruganathan, S., Joty, S., Ouzzani, M., Tang, N.: Deeper–deep entity resolution. arXiv preprint arXiv:1710.00597 (2017)
6. Ebraheem, M., Thirumuruganathan, S., Joty, S., Ouzzani, M., Tang, N.: Distributed representations of tuples for entity resolution. vol. 11, pp. 1454–1467. VLDB Endowment (2018)
7. IFLA Study Group on the Functional Requirements for Bibliographic Records : Functional requirements for bibliographic records - final report (Feb 2009), `https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf`
8. Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., Schwab, D.: Flaubert: Unsupervised language model pretraining for french. arXiv preprint arXiv:1912.05372 (2019)
9. Li, Y., Li, J., Suhara, Y., Doan, A., Tan, W.C.: Deep entity matching with pretrained language models. arXiv preprint arXiv:2004.00584 (2020)
10. Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de la Clergerie, É.V., Seddah, D., Sagot, B.: Camembert: a tasty french language model. arXiv preprint arXiv:1911.03894 (2019)
11. Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., Deep, R., Arcaute, E., Raghavendra, V.: Deep learning for entity matching: A design space exploration. In: Proceedings of the 2018 International Conference on Management of Data. pp. 19–34. VLDB Endowment (2018)
12. Papadakis, G., Ioannou, E., Palpanas, T.: Entity resolution: Past, present and yet-to-come. In: EDBT. pp. 647–650 (2020)
13. Riva, P., Le Bœuf, P., Žumer, M.: Ifla library reference model. A Conceptual Model for Bibliographic Information. Hg. v. IFLA International Federation of Library Associations and institutions. Online verfügbar unter https://www. ifla. org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712. pdf (2017)
14. Suchanek, F.M., Abiteboul, S., Senellart, P.: Paris: Probabilistic alignment of relations, instances, and schema. arXiv preprint arXiv:1111.7164 (2011)