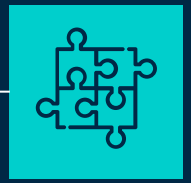
Masked Language Model Entity Matching for Cultural Heritage Data

Dominique Piché, Amal Zouaq, Michel Gagnon and Ludovic Font

École Polytechnique de Montréal

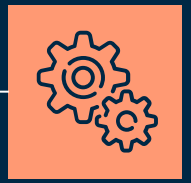
TABLE OF CONTENTS



01

Introduction

1. Entity Matching
2. Neural matching



02

Experimental setup

1. Data overview
2. MCC KB model
3. Methodology

03

Analysis

1. Results
2. Limits
3. Further work

Entity Matching

Problem:

How to determine entries describing two same real-world entities, in the absence of unique identifiers and with unclean data?

Matching Literary Metadata

Challenges

-> Though global identifiers exist (ISNI, VIAF, ISBN), many datasets do not include them systematically and rely on internal identifiers

-> Data quality and models are highly variable:

- Variations in chosen models: MARC21, Onix, in-house solutions
- Variations in encoding conventions: Surname, First Name vs. First Name Surname
- Mislabeled values, erroneous characters
- Textual description fields

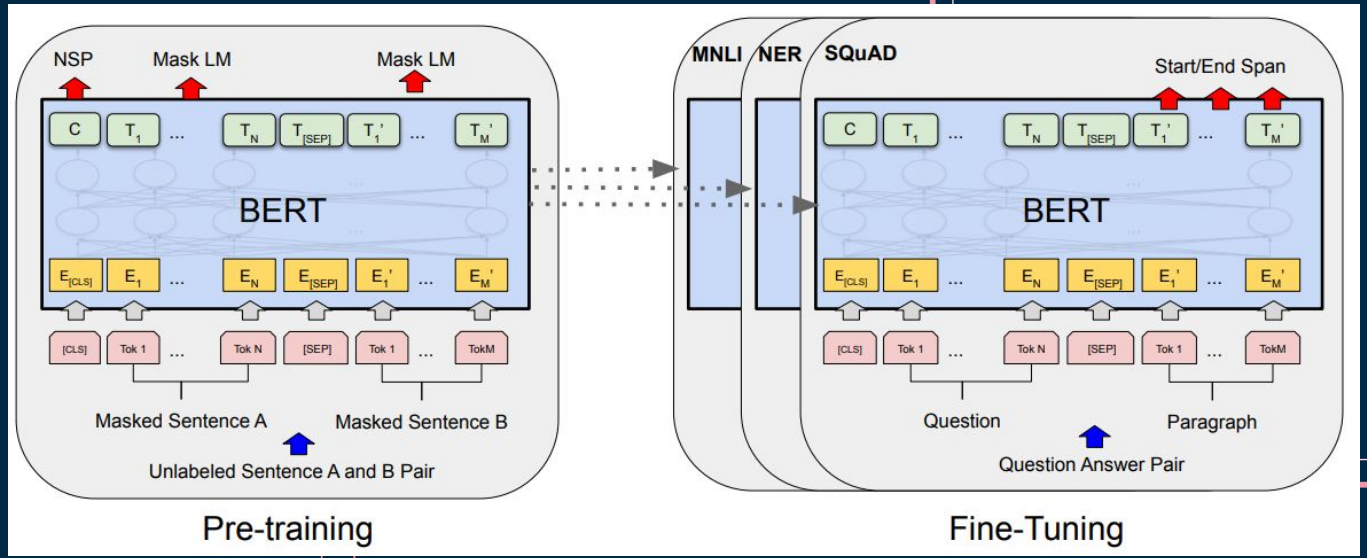
Examples of non-standardized fields

Titles:
-> Long cri dans la nuit -un
-> Guide prat.vin d'italie
-> Ultra Quiz QI - NÂ° 2
-> Les maîtres du Pentacle - Tome 2

Names:
-> BARBARA PH.D DEANGELIS
-> Collectif
-> J. A. Gonzalez
-> Couture, Yvon H.

BERT

Bidirectional Encoder Representations from Transformers



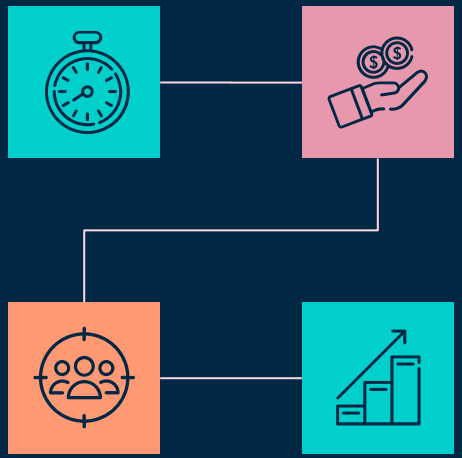
https://arxiv.org/pdf/1810.04805.pdf

Neural Matching

Ditto: using Transformers for Entity Matching

Pre-trained Transformer

DistilBERT, RoBERTa, ALBERT



Serialization

The COL column VAL value format

$$\text{serialize}(e) ::= [\text{COL}]\ \text{attr}_1\ [\text{VAL}]\ \text{val}_1 \ldots [\text{COL}]\ \text{attr}_k\ [\text{VAL}]\ \text{val}_k,$$

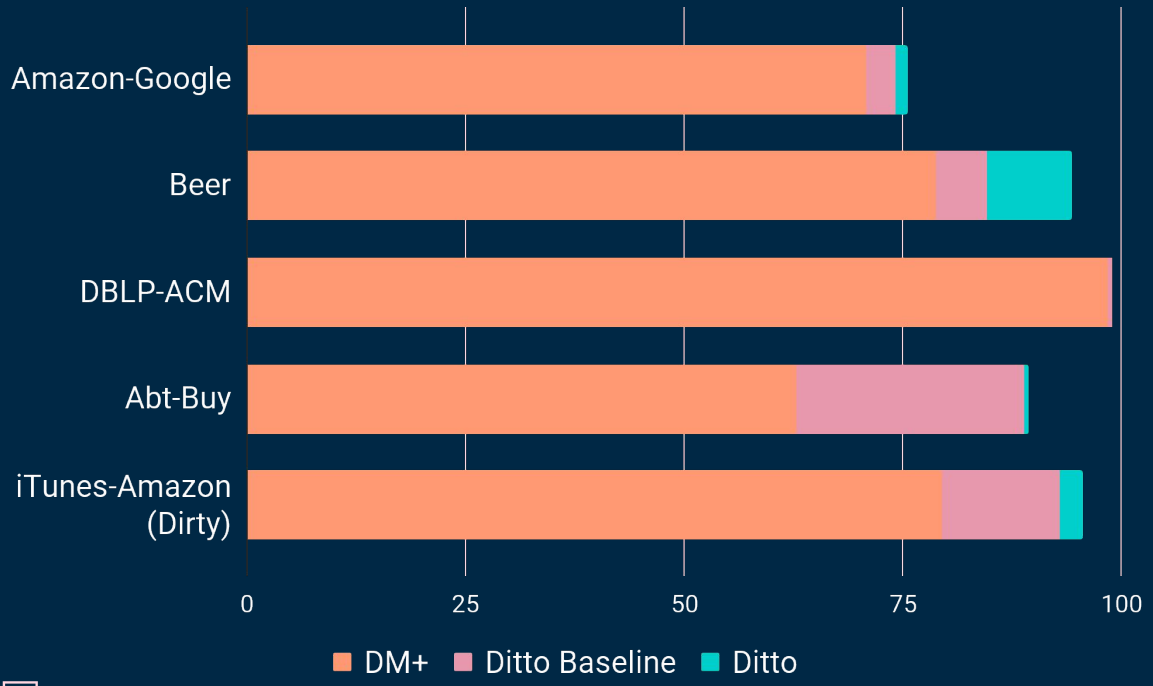
Sequence-pair classification problem

Binary classification using Fully Connected and Softmax layers

Finetuning

Labeled sets of serialized entries

ER Magellan Results



Data Sources

BAnQ
Bibliothèque et Archives nationales du Québec
-> Legal Deposit for Québec

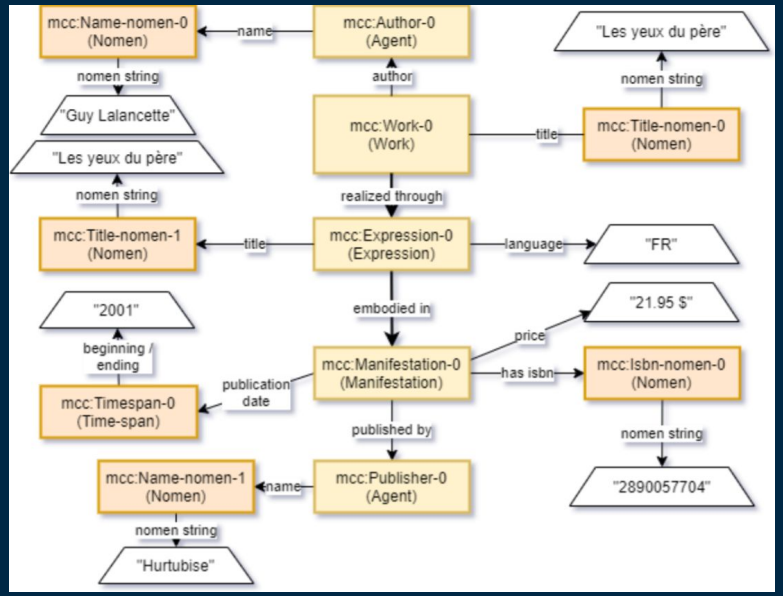
ADP
Book distributor

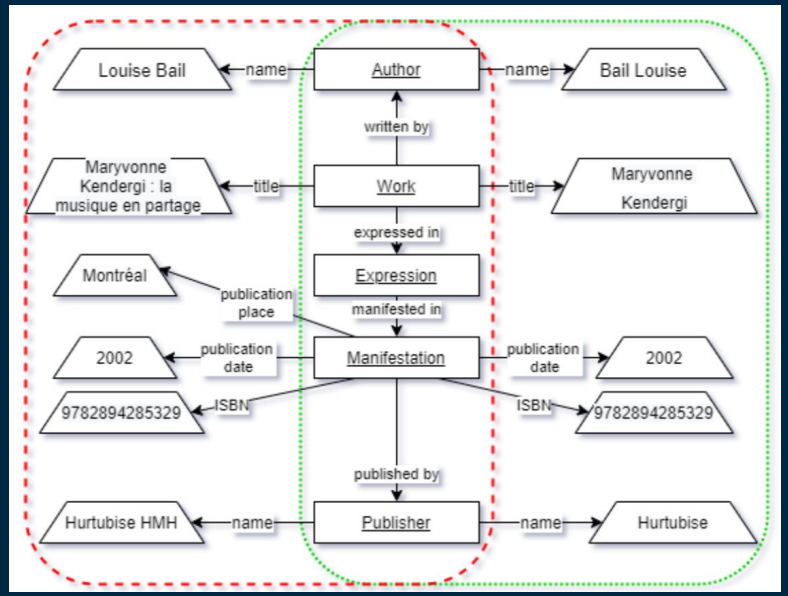
l'Île
Union des écrivaines et écrivains québécois
-> Québec's Writers Union

Hurtubise
Book publisher

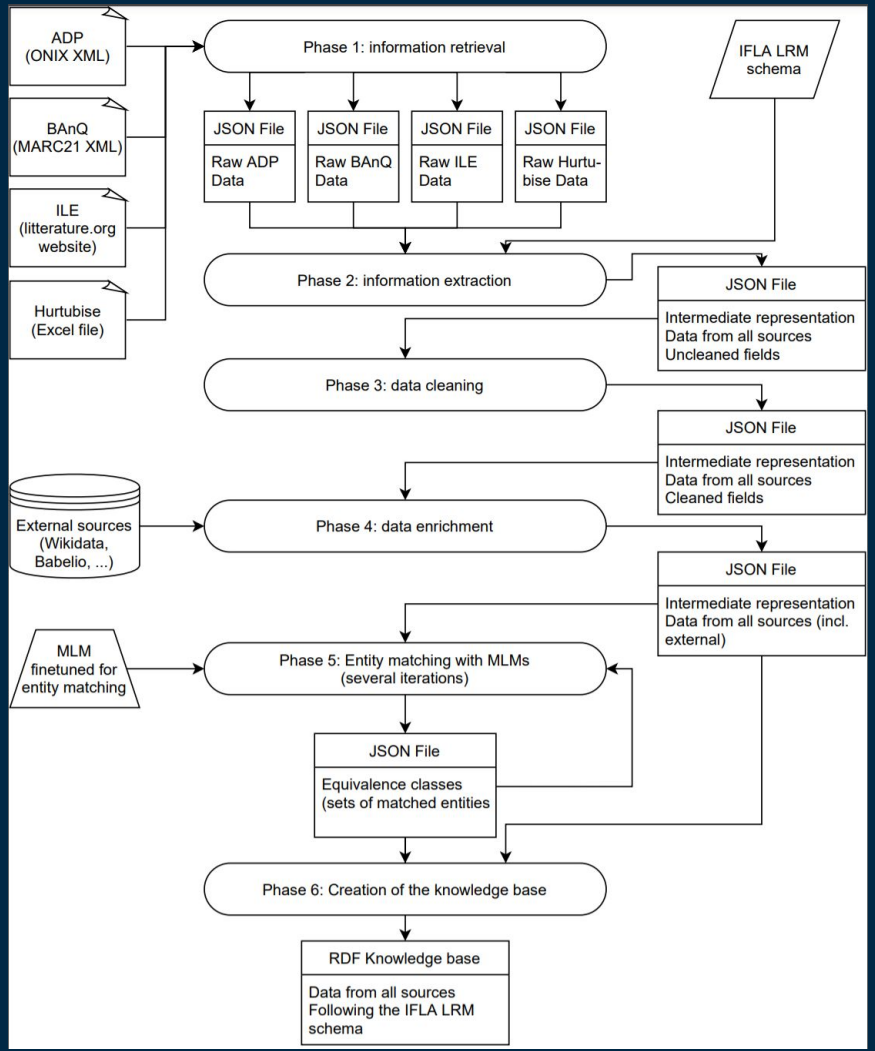
					Integrated knowledge base	
Entity type	BAnQ	ADP	Hurtubise	l'Île	Before alignment	After alignment
Work	50,140	13,707	1,012	21,726	86,585	74,764
Expression	58,476	14,772	1,056	26,203	100,507	87,007
Manifestation	79,687	15,399	2,505	27,662	125,283	109,240
Author	16,960	6,235	636	1,766	25,597	22,692
Publisher	3,112	38	1	2,315	5,466	3,921
Total	208,375	50,151	5,210	79,702	343,438	297,624

MCC Knowledge Base Model





Pipeline


ADP (ONIX XML)
BAnQ (MARC21 XML)
ILE (litterature.org website)
Hurtubise (Excel file)
Phase 1: information retrieval
IFLA LRM schema
JSON File
Raw ADP Data
JSON File
Raw BAnQ Data
JSON File
Raw ILE Data
JSON File
Raw Hurtu-bise Data
Phase 2: information extraction
JSON File
Intermediate representation
Data from all sources
Uncleaned fields
Phase 3: data cleaning
JSON File
Intermediate representation
Data from all sources
Cleaned fields
External sources (Wikidata, Babelio, ...)
Phase 4: data enrichment
JSON File
Intermediate representation
Data from all sources (incl. external)
MLM finetuned for entity matching
Phase 5: Entity matching with MLMs (several iterations)
JSON File
Equivalence classes (sets of matched entities
Phase 6: Creation of the knowledge base
RDF Knowledge base
Data from all sources
Following the IFLA LRM schema

Methodology

-> Entity matching with BERT-like models:

- French MLMs: CamemBERT, RoBERTA
- Entity types: Authors and Works
- Data cleaning: cleaned and dirty
- Data model: original and LRM
- Attribute labeling: unlabeled, labeled, labeled and tagged
- Separate models for Authors and Works vs. Joint Author-Work matcher

Entity	Rule	Match	Description
Work	1	Title (0.90) & Subtitle (0.90)	Similar titles and subtitles
	2	Title (0.90) & Author name (0.95)	Similar titles and Author names
	3	Title (0.90) & Publication year (1)	Similar titles and same publication years
Author	1	Name (1)	With sources without Author's Works
	2	Name (0.95) & Birth year (1)	Name and birth year match
	3	ISNI (1)	ISNI is a unique URI
	4	ILE URI (1)	ILE URI is a unique URI
	5	Name (0.95) & Work Title (0.80)	Similar names and one similarly titled Work
Publisher	1	Authority list (0.95)	Name close matched to authority list
	2	Publisher name (0.95)	Publishers with similar names

Heuristic baseline rules

-> Comparison with Heuristic baseline

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

Size of training and evaluation sets

Input formats

1: Tagged, labeled and cleaned, with LRM-based model attribute names

2: Tagged, labeled and dirty, with original model attribute names

3: Labeled and dirty, with original model attribute names

4: Unlabeled and dirty, without attribute names

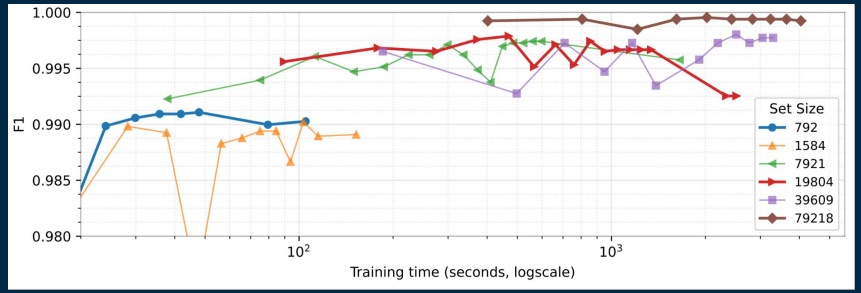
#		Value
1	Seq 1 :	**[C] t [V]** Être un héros **[C] e [V]** La Courte échelle **[C] st [V]** des histoires de gars **[C] lp [V]** Montréal **[C] np [V]** 218
	Seq 2 :	**[C] t [V]** Être un héros **[C] ap [V]** 2011 **[C] a [V]** Simon Boulerice **[C] e [V]** La Courte échelle **[C] st [V]** des histoires 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

Results

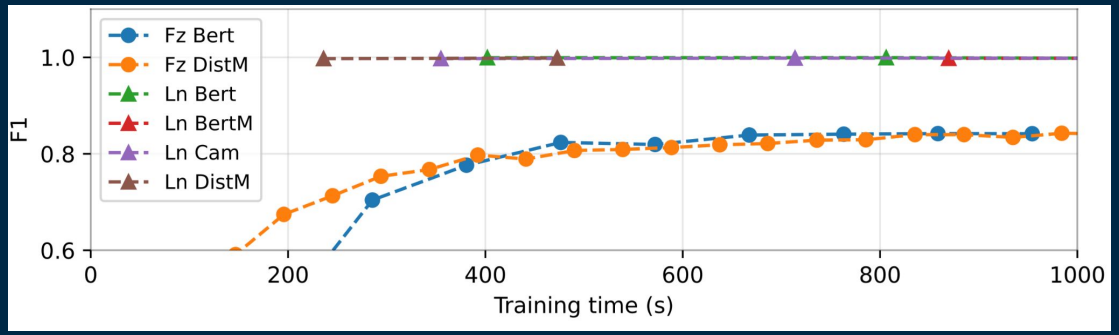
A. **Best matcher** versus **baseline**			**Evaluation on test set**			
Model Type	Peak Ep.	Entity	F1	Recall	Precision	Accuracy
Rule-based		Author	0.9955	0.9965	0.9945	0.9970
		Work	0.9119	0.8407	0.9962	0.9458
CamemBERT	4	Author	**0.9986**	**0.9979**	**0.9993**	**0.9991**
	4	Work	**0.9978**	**0.9968**	**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**

Supplemental Results

Impact of Training set size



Impact of MLM weight shifting



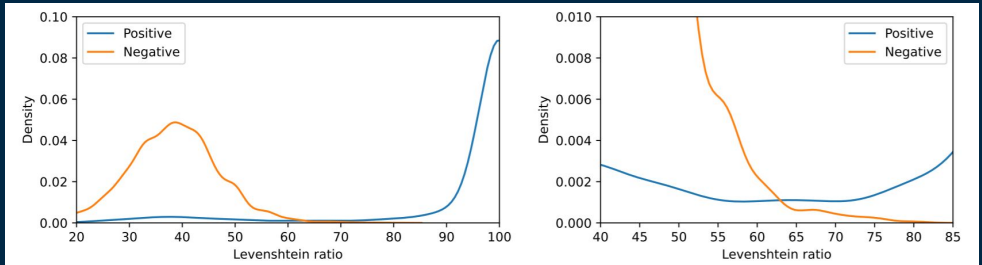
Limits

-> Character limits

-> Negative example generation

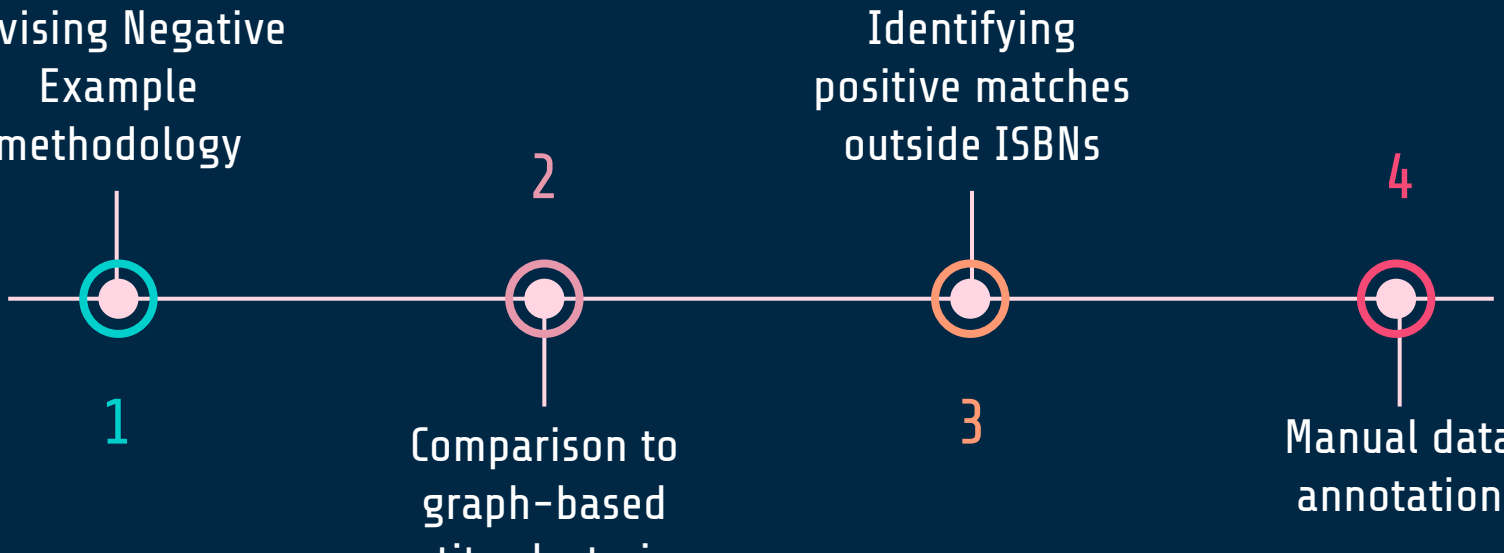
-> Exploiting graph structure

-> Use of ISBNs for positive example generation



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

Further Work



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)

Do you have any questions?

THANK YOU