From teaching books to educational videos and vice versa: a cross-media content retrieval experience

From teaching books to educational videos and vice versa: a cross-media content retrieval experience

Luca Cagliero, Lorenzo Canale, Laura Farinetti
Politecnico di Torino
Torino, Italy
(luca.cagliero, lorenzo.canale, laura.farinetti)@polito.it

***Abstract*—Due to the rapid growth of multimedia data and the diffusion of remote and mixed learning, teaching sessions are becoming more and more multi-modal. To deepen the knowledge of specific topics, learners can be interested in retrieving educational videos that complement the textual content of teaching books. However, retrieving educational videos can be particularly challenging when there is a lack of metadata information. To tackle the aforesaid issue, this paper explores the joint use of Deep Learning and Natural Language Processing techniques to retrieve cross-media educational resources (i.e., from text snippets to videos and vice versa). It applies NLP techniques to both the audio transcript of the videos and to the text snippets in the books in order to quantify the semantic relationships between pairs of educational resources of different media types. Then, it trains a Deep Learning model on top of the NLP-based features. The probabilities returned by the Deep Learning model are used to rank the candidate resources based on their relevance to a given query. The results achieved on a real collection of educational multimodal data show that the proposed approach performs better than state-of-the-art solutions. Furthermore, a preliminary attempt to apply the same approach to address a similar retrieval task (i.e., from text to image and vice versa) has shown promising results.**

***Index Terms*—Learning analytics, cross-media retrieval, educational data mining, deep learning**

I. Introduction

The current COVID-19 pandemic event dramatically changed higher education practice [1], [2] and its impact will likely be long-lasting presenting both challenges and opportunities [3]. The situation forced most universities worldwide to move to a completely remote or a mixed learning modality, where teaching sessions generally are streamed and also recorded to be available to students that are not able to attend them in real time. This brought to a huge increase of available video educational material, which will be a potentially valuable asset for learners in the future [4], [5], provided that it is easily searchable and connected with other educational material.

The need for cross-media retrieval solutions, capable of retrieving different media types such as videos and images, is therefore evident. The possibility to create automatic links between different types of educational material has been enhanced by recent advances in Natural Language Processing and Deep Learning [6]. The former techniques provide semantically rich text descriptions, whereas the latter support the learning of accurate inference models from large sets of features, which provide latent descriptions of the multi-modal resources [7].

Thanks to the tight integration with semantic web models, cross-media retrieval systems can foster the learner's capability to create knowledge networks according to the constructivist theory of learning [8]. Besides, mixing educational resources of different media types can enhance the learning engagement of the students [4], [9] and increase their ability to focus on learning outcomes.

The present work focuses on supporting learners who are exploring either a fragment of a teaching book or an educational video to retrieve pertinent resources of the other media type, i.e., from text to video or vice versa. A major challenge in the specific task is that educational videos produced as the result of online video-lectures are typically far from being high-quality "educational pills", such as the TED model [10]. Furthermore, since videos are usually not produced by following a standard design process such as MOOCs [11], [12], they are often not annotated with semantically rich metadata. This hinders the use of querying services available in the most common digital libraries, thus calling for alternative cross-media retrieval solutions tailored to the learning context.

This paper presents a new Learning Analytics application consisting of a cross-media content retrieval system specifically designed for handling teaching books and educational videos with limited or absent annotations. To overcome the lack of metadata information, it applies NLP techniques to extract semantically rich text descriptions from both book text snippets and audio transcripts of the videos. The extracted cross-media knowledge is then collected into a dataset whose NLP-based features summarize the level of similarity between pairs of snippets and videos from various viewpoints. Next, a Deep Learning model is trained on the prepared dataset in order to automatically infer the presence or absence of a link between a given pair of video and book text snippet. The output probabilities are then exploited to rank the retrieved resources based on the input query.

The proposed system outperformed existing cross-media retrieval methods on a real educational dataset. Furthermore, it achieved promising preliminary results in a similar retrieval task, where images are retrieved instead of educational videos.

The paper is organized as follows. Section II presents the related works. Section III formalizes the cross-media retrieval task. Section IV describes the methodology. Section V reports

the experimental results, whereas Section VI draws the conclusions of the present work.

II. Related works

This work is an application of cross-media retrieval techniques to educational books and videos. Hence, hereafter we will discuss the position of the work in both the information retrieval and learning analytics domains.

a) Cross-media retrieval: Cross-media retrieval is a well-known Information Retrieval task, where the goal is to allow end-users to submit queries of a particular media type (e.g., text) and to retrieve pertinent results of a different type (e.g., videos) [13]. The mainstream in cross-media retrieval is to learn a common space where multi-modal descriptions of semantically related resources can be effectively and efficiently compared with each other. According to [6], existing approaches can be classified as:

- *Graph-based methods*, where the set of pairwise correlations between cross-media resources is modelled as a weighted graph [14], [15]. According to the strategy used to extract the resources that are most relevant to a given query, they can be further partitioned into *graph regulations strategies* (e.g., [16], [17]) and *neighbor analysis methods* (e.g., [18], [19]).
- *Learning to Rank methods*, which reformulate the retrieval task as a ranking optimization problem by using ranking information as training data (e.g., JRL [20], Bi-CMSRM [21], CLM^{2}R [22]).
- *Hashing methods*, which generate the hash codes for more than one media type and then project cross-media data into a common Hamming space (e.g., [23], [24]).
- *Deep Learning (DL) methods*, which train Deep Neural Network models to mine complex relationships among cross-media content (e.g., [25], [26]). A recent survey of DL-based methods is given in [27].
- *Natural Language Processing (NLP)-based methods*, which processes the natural language expressed in textual form by trying to embed the information extracted in the different modalities (e.g., [28], [29]).

The present study introduces an hybrid strategy that combines DL with NLP. It focuses on automatically retrieving cross-media content in a educational context where we have limited access to descriptive metadata (or they are almost missing).

b) Content retrieval in Learning Analytics: Several attempts to effectively and efficiently retrieve educational resources have previously been made. For example, the works presented in [30], [31] proposed different strategies to index and search video lectures. The mainly addressed challenges are video segmentation and metadata extraction from OCR content. Unlike [30], [31], we address content retrieval in a cross-media scenario. In [32] the authors exploited both speech and video information to automatically retrieve video lectures. They extract metadata from video content by automatically detecting slide text. However, multi-modal data are gathered from the same resource. Furthermore, in the proposed pipeline supervised learning is applied to perform slide segmentation, whereas our approach applies Deep Learning models to derive the rank of the retrieved videos.

In parallel, some efforts to understand the importance of handling multi-modal content in education have been made. For example, in [33]–[35] the authors analyzed YouTube videos in order to derive their cognitive value and practical usefulness. The work presented in [**?**] focused on extracting and recommending textual summaries of teaching books, whereas in [36] and [37] the authors investigated the educational role of augmented reality and social platforms, respectively. The current work proposes a new Deep Learning application focused on enhancing the accessibility of multi-modal resources.

III. Problem statement

The aim of the present work is to ease the retrieval of related educational resources of different media types, i.e., from teaching books to educational videos and vice versa.

Let $\mathcal{B}$ and $\mathcal{V}$ be the set of considered books and videos, respectively. For our purposes, we split the content of a book $b \in \mathcal{B}$ into a set of text snippets TS=$\{ts_1, ts_2, \ldots, ts_{|TS|}\}$. Text splitting is commonly based on the book structure. For example, a text snippet can be mapped to a specific book chapter, subsection, or paragraph.

Each video $v \in \mathcal{V}$ can be enriched with a set $\mathcal{M}_v$ of metadata information, whereas each metadata resource $m_v \in \mathcal{M}_v$ has a type (e.g., title, author, category, duration, language, audio transcript) and takes value for each video. Hereafter, we will mainly focus on the audio transcript of the video because in the educational context video-lectures are often poorly annotated.

Let $sim(ts,v)$ be the cross-media similarity between a text snippet ts and a video $v \in \mathcal{V}$. Without any loss of generality, let $sim(ts,v)$ be a boolean function denoting whether the resource pair is pertinent (1) or not (0). Our preliminary goal is to learn a supervised model that is able to accurately estimate the probabilities $p(sim(ts,v)$=1) and $p(sim(ts,v)$=0) for each pair $\langle ts, v\rangle$ such that $sim(ts,v)$ is unknown. Probability estimates rely on an initial sample of manually labeled pairs (i.e., the training set). For each resource pair the training set will incorporate the syntactical and semantic similarities between the textual content in ts and the video metadata $\mathcal{M}_v$.

Given a query $Q \in \mathcal{V} \cup TS$ consisting of an arbitrary resource, the top-k cross-media retrieval task entails retrieving relevant instances of the other media type [6]. For example, given a text snippet in a book we retrieve the top-k most relevant videos or given a video we retrieve the top-k snippets.

Let $\mathbf{R}_Q^{(k)}$ the size-k vector of retrieved resources for Q. The aforesaid task can be formulated as follows:

$$\arg_{\mathbf{R}_Q^{(k)}} \max \left(p\Big(sim(Q, \mathbf{x}) = 1 \Big) \right)$$

where $\mathbf{x}$=v if $Q \in TS$ (i.e., retrieve video from text) whereas $\mathbf{x}$=ts otherwise, i.e., $Q \in V$ (i.e., retrieve text from video).

IV. Cross-media retrieval: the proposed methodology

The cross-media retrieval approach relies on Deep Learning model trained on NLP features. The key idea is to analyze the syntactical and semantic relationships between the content of the text snippets in the teaching book and the metadata values extracted from the educational video. We train a classification model that predicts whether the pairwise association between text snippets and videos are appropriate based on the similarities between the NLP descriptions of the two resources.

To capture the semantics behind the analyzed text, we rely on the Wikidata Knowledge Base (KB). The ability of Wikidata to accurately describe and retrieve domain-specific concepts enables the effective characterization and mapping of the cross-media content [38]. Furthermore, the hierarchical organization of the KB concepts is instrumental in enriching the concepts recognized from the resource descriptors thus improving the likelihood to find correct matches between pairs of cross-media resources.

The proposed approach consists of the following steps, which are also summarized in Figure 1:

1) *Text extraction*: it focuses on processing the textual content of both the teaching book and the metadata of the educational videos in order to produce comparable resource descriptions (see Section IV-A).
2) *Named Entity Linking*: it entails extracting from the textual descriptions semantically relevant KB entities (see Section IV-B).
3) *Feature engineering*: it addresses the generation of a cross-media feature set which incorporates the main syntactic and semantic relationships between pairs of cross-media resources (see Section IV-C).
4) *DNN-based inference and ranking*: it focuses on inferring the cross-media links between pairs of resources of different media types based on the previously generated features. Then, given a resource of a specific media type in the input query, it explores the inference outcomes in order to retrieve the top-k most pertinent resources in the other media type (see Section IV-D).

A more detailed description of each step follows.

A. Text extraction

We build a corpus that consists of pairs $\langle ts, v \rangle$ of textual snippets ts and video v described by the corresponding metadata M_v.

To split plain text into snippets, we detect the physical structure of the books using the *PyPDF2* library[1]. For each textual snippets we extract the corresponding plain text by using the *ConvertApi* service[2], whose exposed Application Programming Interfaces support multiple data types.

To extract the audio transcript of the educational videos, we exploit *Cloud Natural Language APIs*[3].

[1] http://mstamy2.github.io/PyPDF2/ (latest access: January 2021)
[2] https://www.convertapi.com/pdf-to-txt (latest access: January 2021)
[3] https://bit.ly/3igr1x2 (latest access: January 2021)

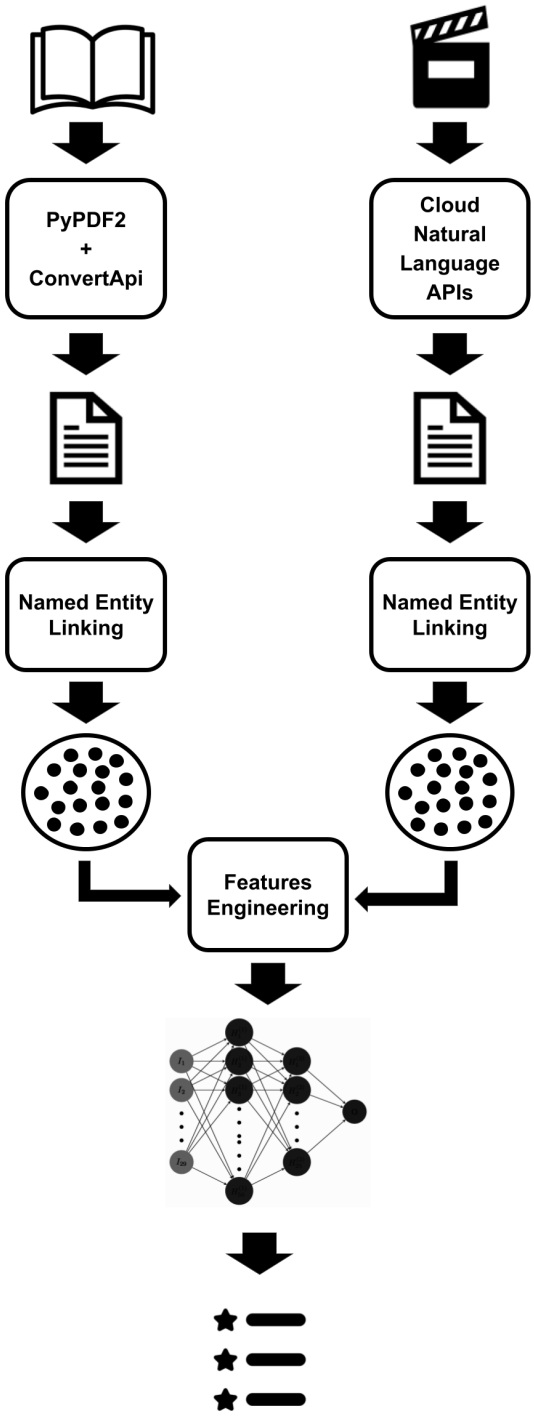

Fig. 1: Sketch of the proposed methodology.

To avoid introducing a bias in the next learning phase, the textual context retrieved from both media types is cleaned up using ad hoc data cleaning procedures (e.g., we automatically remove time codes from the audio transcript).

B. Named Entity Linking

This step aims at splitting the raw text in separate units, commonly denoted as *tokens*, and then searching for a mapping to real-world (abstract or existing) objects such as persons, locations, organizations, products [7]. The recognized objects are formally described as *named entities* in a shared knowledge base and can be referenced with the corresponding Uniform Resource Identifier (URI).

We rely on the *Wikidata*[4] knowledge base [38] since it achieved higher-quality standards compared to alternative solutions [39]. Notice that Wikidata content curation relies on a voluntary basis, requires community approval prior to adding new content, and supports data ingestion from external data sources.

Wikidata organizes the large set of available entities into complex hierarchies [40]. The above-mentioned structure is not only suitable for effectively tackling the standard Named Entity Recognition task[5] but it also provides end-users with a multi-resolution description of the underlying concepts, which is particularly useful, in our context, for differentiating educational entities in cross-media content.

To extract Wikidata named entities from plain text we apply the following extractors: *TextRazor*[6], *Babelfy*[7] and *Google Cloud Speech API*[8]. Since the recognized entities are not necessarily part of the Wikidata knowledge base, whenever necessary we conveniently merge all the intermediate results provided by each single extractor and possibly relative to different knowledge bases. More specifically, we exploit the semantic cross-link named `owl:sameAs` in order to map the external entities to the official Wikidata ones.

C. Feature engineering

We quantify the syntactical and semantic relationships between the textual description of the cross-media resources and store them in a multivariate dataset, which will be used later on to automatically infer cross-media resource links.

Let E_{ts} and E_v be the set of KB entities associated with the text snippet ts and video v. For both entity sets we first distinguish instances from other entity types by querying the Wikidata Sparql endpoint as follows.

```
ASK {
  <ENTITY URI> wdt:P31 ?o.
}
```

where `ENTITY URI` indicates the URI of the entity involved, whereas `wdt:P31` is the *instance of* predicate in the KB.

Let I_{ts} and I_v be the entity sets consisting of all the instances in E_{ts} and E_v, respectively. In order to semantically enrich the E_{ts} and E_v contextual descriptions, we query the KB to retrieve the corresponding parents of all the considered entities by using the predicates `wdt:P279` (*subclass of*) and `wdt:P361` (*part of*).

```
SELECT ?o {
  <ENTITY URI> wdt:P279|wdt:P361 ?o.
}
```

Let P_{ts} and P_v be the parent entities related to E_{ts} and E_v, respectively. The extended entity sets E^*_{ts} and E^*_v are obtained by the union of the respective child and parent entities P_{ts} and P_{ts}, i.e., $E^*_{ts} = E_{ts} \cup P_{ts}$, $E^*_{ts} = E_v \cup P_v$.

The feature set is derived from the intersection and union of the entities in P_{ts}, E_{ts}, I_{ts}, E^*_{ts}, P_v, E_v, I_v, E^*_v according to the similarity measures described in [41] and [42]. The considered feature set is summarized in Table I.

D. DNN-based inference and ranking

We train a Deep Neural Network model on the prepared training data in order to support the automatic inference of the cross-media resource links. Specifically, we exploit a fully connected 2-layer Neural Network architecture. The model takes as inputs $I_1, I_2, \ldots, I_{22}$, i.e. the values of the cardinality- and similarity-based features associated with an unlabeled pair of text snippet and video (see Section IV-C). To avoid introducing a bias in the learning phase, feature values were preemptively normalized using a min-max scaler:

$$X_{norm} = 2\left(\frac{X - X_{min}}{X_{max} - X_{min}}\right) - 1$$

Cross-media content retrieval relies on the network outputs (O) produced by taking the queried resource (Q) combined with any candidate resources. The top-k resources in order of decreasing output probability are retrieved.

V. Experimental results

We empirically evaluated the performance of the proposed approach on a real collection of cross-media educational content.

The remainder of this section is organized as follows. Sections V-A and V-B respectively describe the analyzed data collection and the network configuration settings. Sections V-C and V-D enumerate and briefly describe the tested competitors and the evaluation metrics, respectively. Section V-E analyzes the relevance of the NLP-based feature categories enumerated in Table I to accurately predict cross-media resource links. Sections V-F and V-G respectively report the outcomes of the quantitative and qualitative evaluations. Finally, Section V-H presents the preliminary results of a transfer learning experiment in which the proposed method was applied to solve a similar problem in a related domain, i.e., the cross-media text-image task.

[4] http://wikidata.org/ (latest access: January 2021)

[5] Named Entity Recognition (NER) is a information extraction task aimed at seeking named entities mentioned in unstructured text and classifying them according to a predefined categorization.

[6] https://www.textrazor.com/ (latest access: January 2021)

[7] http://babelfy.org/ (latest access: January 2021)

[8] https://bit.ly/3ifyQmv (latest access: January 2021)

TABLE I: Feature set used to describe the $\langle ts,v \rangle$ pairs

Index	Feature	Description
Cardinalities		
1	$\lvert E_{ts} \rvert$	cardinality of set E_{ts}
2	$\lvert E_v \rvert$	cardinality of set E_v
3	$\lvert E_v \cap E_{ts} \rvert$	cardinality of the intersection between E_v and E_{ts}
4	$\lvert E_v \cup E_{ts} \rvert$	cardinality of the union between E_v and E_{ts}
5	$\lvert I_{ts} \rvert$	cardinality of set I_{ts}
6	$\lvert E_v \cap I_{ts} \rvert$	cardinality of the intersection between E_v and I_{ts}
7	$\lvert E_v \cup I_{ts} \rvert$	cardinality of the union between E_v and I_{ts}
8	$\lvert I_v \rvert$	cardinality of set I_v
9	$\lvert E_{ts} \cap I_v \rvert$	cardinality of the intersection between E_{ts} and I_v
10	$\lvert E_{ts} \cup I_v \rvert$	cardinality of the union between E_{ts} and I_v
11	$\lvert I_{ts} \cap I_v \rvert$	cardinality of the intersection between I_{ts} and I_v
12	$\lvert I_{ts} \cup I_v \rvert$	cardinality of the union between I_{ts} and I_v
13	$\lvert E_t^* \cap E_v^* \rvert$	cardinality of the intersection between E_t^* and E_v^*
14	$\lvert E_t^* \cup E_v^* \rvert$	cardinality of the union between E_t^* and E_v^*
Similarities		
15	$\frac{\lvert E_v \cap E_{ts} \rvert}{max(\lvert E_v \rvert, \lvert E_{ts} \rvert)}$	normalized weighted intersection between E_v and E_{ts}
16	$\frac{\lvert E_v \cap E_{ts} \rvert}{min(\lvert E_v \rvert, \lvert E_{ts} \rvert)}$	overlap coefficient between E_v and E_{ts}
17	$\frac{\lvert E_v \cap E_{ts} \rvert}{\lvert E_v \cup E_{ts} \rvert}$	Jaccard similarity between E_v and E_{ts}
18	$\frac{\lvert E_v \cap I_{ts} \rvert}{max(\lvert E_v \rvert, \lvert I_{ts} \rvert)}$	normalized weighted intersection between E_v and I_{ts}
19	$\frac{\lvert E_v \cap I_{ts} \rvert}{min(\lvert E_v \rvert, \lvert I_{ts} \rvert)}$	overlap coefficient between E_v and I_{ts}
20	$\frac{\lvert E_v \cap I_{ts} \rvert}{\lvert E_v \cup I_{ts} \rvert}$	Jaccard similarity between E_v and I_{ts}
21	$\frac{\lvert E_{ts} \cap I_v \rvert}{max(\lvert E_{ts} \rvert, \lvert I_v \rvert)}$	normalized weighted intersection between E_{ts} and I_v
22	$\frac{\lvert E_{ts} \cap I_v \rvert}{min(\lvert E_{ts} \rvert, \lvert I_v \rvert)}$	overlap coefficient between E_{ts} and I_v
23	$\frac{\lvert E_{ts} \cap I_v \rvert}{\lvert E_{ts} \cup I_v \rvert}$	Jaccard similarity between E_{ts} and I_v
24	$\frac{\lvert I_{ts} \cap I_v \rvert}{max(\lvert I_{ts} \rvert, \lvert I_v \rvert)}$	normalized weighted intersection between I_{ts} and I_v
25	$\frac{\lvert I_{ts} \cap I_v \rvert}{min(\lvert I_{ts} \rvert, \lvert I_v \rvert)}$	overlap coefficient between I_{ts} and I_v
26	$\frac{\lvert I_{ts} \cap I_v \rvert}{\lvert I_{ts} \cup I_v \rvert}$	Jaccard similarity between I_{ts} and I_v
27	$\frac{\lvert E_{ts}^* \cap E_v^* \rvert}{max(\lvert E_{ts}^* \rvert, \lvert E_v^* \rvert)}$	normalized weighted intersection between E_{ts}^* and E_v^*
28	$\frac{\lvert E_{ts}^* \cap E_v^* \rvert}{min(\lvert E_{ts}^* \rvert, \lvert E_v^* \rvert)}$	overlap coefficient between E_{ts}^* and E_v^*
29	$\frac{\lvert E_{ts}^* \cap E_v^* \rvert}{\lvert E_{ts}^* \cup E_v^* \rvert}$	Jaccard similarity between E_{ts}^* and E_v^*
Target		
class	$sim(ts,v)$	1=relevant, 0=not relevant

A. Educational dataset

To build the education dataset, we chose the electronic versions of ten open teaching books related to computer science and a corresponding set of educational videos presenting specific topics covered in the books (e.g., the use of the Python language in data analytics). We split books into chapters (separated by titles) and selected the ten most representative chapters per book. The key book and video characteristics are summarized in Table II. Text snippets and videos are rather diversified in length/duration.

TABLE II: Dataset statistics.

Property	Value		
	Min	Max	Avg
No. of snippets per book	51	391	213
Snippets length (word count)	2	34359	2959.6
No. of relevant videos per snippet	1	10	4.7
Video length	2min	1h21min	21min
Title length (word count)	1	9	3.2
Transcript length (word count)	1200	20243	2627.2

B. Network configuration settings

To train the Deep Learning models we exploited the Python-based network implementations available in Keras [43].

To evaluate classification performance, we first split the prepared dataset into train, validation and test sets. Next, we trained the model on the training set by performing a grid search on the validation set in order to find the network setup achieving the least loss value. A summary of the considered hyper-parameter values is given in Table III.

TABLE III: Hyper-parameter values used for the grid search.

Hyperparameter	Set of possible values
activation for middle layers	selu, relu, elu, sigmoid
loss function	cosine similarity, mse, binary cross entropy
gradient descent	adam
dropout for middle layers	0.05, 0.10, 0.15, 0.20, 0.25, 0.30
batch size	10, 50, 100, 250, 500
number of middle layers	1,2,3,5,10
units per layer	10, 25, 50,100

C. Competitors

We compared our methodology against state-of-the-art algorithms. Specifically, based on the results reported in [6], we picked the two best performing algorithms for cross-media entity retrieval from video to text and form text to video, i.e., JRL [20] and JGRHML [16]. The former approach embeds the various information sources into a unified space based on Deep NLP models architectures. The latter proposes a methodology based on graph regularization. Furthermore, we tested also the following three baseline methods relying on the established contextualized BERT embeddings [44]:

- **B-Title**: BERT Similarity with the video title.
- **B-Trans**: BERT Similarity with the video transcript.

- **B-Title+Trans**: BERT Similarity with the video title+transcript.

All the aforesaid methods compute the similarity score, in the latent embedding space, between the textual content available in the two media types. Concerning the video metadata, the former baseline method (*B-Title*) focuses on the video title, the baseline strategy named *B-Trans* on the audio transcript, whereas the latter (*B-Title+Trans*) on both title and audio transcript.

D. Evaluation metrics

In compliance with [6], we evaluated the performance of cross-media retrieval systems in terms of Mean Average Precision (MAP) [45]. Given an input query, the Average Precision (AP) is computed as follows:

$$AP = \frac{\sum_{n}^{k=1} (P(k) \times rel(k))}{num.\ of\ retrieved\ relevant\ resources}$$

where $P(k)$ is the precision at k, whereas $rel(k)$ is an indicator function that takes value one if the retrieved resource at rank k is relevant, zero otherwise.

The MAP score is the average AP over all the performed queries.

E. Feature importance

To explain the reasons behind DNN predictions, we apply a state-of-the-art approach to eXplainable Artificial Intelligence, namely SHapley Additive exPlanations (SHAP) [46]. SHAP provides a visual interpretation of the impact of the feature values taken by a video text pair on the similarities estimated by the Neural Network model. Figure 2 shows the ten most important features and the related SHAP relevance values.

The achieved results show that the most discriminating features are the cardinality of the entity intersection and the overlap coefficient between the two entity sets. Both features do not take into account highly relevant entities in a single media type that have not found in the other media. The reason is that, in most cases, the two separate entity sets contain also entities that not related to the specific educational sub-domain. Hence, considering the entity intersection allows to preserve the quality of the cross-media resource links. Furthermore, the Jaccard similarity score is negatively influenced by not matching entries. The latter feature type gains importance while considering the instances sets I_{ts} and I_v, which are filtered version of E_{ts} and E_v, respectively. The expanded sets E^*_{ts} and E^*_v are clearly less important than the other features. Nevertheless, they appeared in the feature top ten (see, for instance, $\frac{|E^*_{ts} \cap E^*_v|}{max(|E^*_{ts}|,|E^*_v|)}$ and $\frac{|E^*_{ts} \cap E^*_v|}{min(|E^*_{ts}|,|E^*_v|)}$).

F. Quantitative results

Results reported in Table IV show that the proposed method outperforms the competitors in terms of MAP scores. Specifically, its MAP performance is two orders of magnitude higher than the baseline methods and around 40% higher than that of the best performing competitor (i.e., JGRHML [16]).

TABLE IV: Results achieved on the educational dataset.

Methodology	Task	MAP score
Our approach	Text ->Video	**0.527**
	Video ->Text	**0.532**
JGRHML	Text ->Video	0.321
	Video ->Text	0.29
JRL	Text ->Video	0.191
	Video ->Text	0.11
BERT with video title+transcript	Text ->Video	0.011
	Video ->Text	0.008
BERT with video transcript	Text ->Video	0.009
	Video ->Text	0.007
BERT with video title	Text ->Video	0.004
	Video ->Text	0.001

G. Qualitative results

Tables V, VI and VII show three examples of video shortlists retrieved by performing different queries. Each query refers to a chapter selected from a different educational book[9]. The correctly retrieved video identifiers are written in boldface, whereas the ground truth is specified in column *Matching video id*. The precision at rank k ($1 \leq k \leq 3$), i.e., the ability to correctly retrieve resources in the top-k rank, was superior with respect to all the other tested methods.

H. Applicability of the proposed method to similar contexts: preliminary results on the text-image retrieval task

We also assessed the portability of the proposed methodology towards similar retrieval tasks. Specifically, we made a preliminary attempt to apply the proposed method to solve the text-image cross-media retrieval task on a Wikipedia benchmark dataset [47][10]. The steps of the methodology are summarized in Figure 3. The aim is to perform preliminary assessment of the generality of the proposed approach, which is tailored to a specific use case, by testing it in a similar scenario.

The benchmark dataset used to carry out the experiments was already split into training and testing data. For our purposes, we further split train data in order to build the validation set.

To extract Wikidata entities from images, we exploited the *Cloud Vision API*[11], which detects named entities from images using Wikipedia as KB. The corresponding Wikidata entities were derived using the *Wikibase Api*[12].

The grid search on the validation set selected the following hyperparameter values: *activation for middle layers* = *selu*, *loss function* = *mse*, *dropout for middle layers* = 0.10, *batch size* = 250, *number of middle layers* = 2, *units per layer 1* = 50, *units per layer 2* = 10.

Table VIII compares the performance (in terms of MAP scores on test data) achieved by our approach with that of the state-of-the-art algorithms described in [6].

[9]Queries and results were anonymized for double-blind review.

[10]http://www.svcl.ucsd.edu/projects/crossmodal/ (latest access: January 2021)

[11]https://cloud.google.com/vision/docs (latest access: January 2021)

[12]https://wikibase-api.readthedocs.io/en/latest/ (latest access: January 2021)

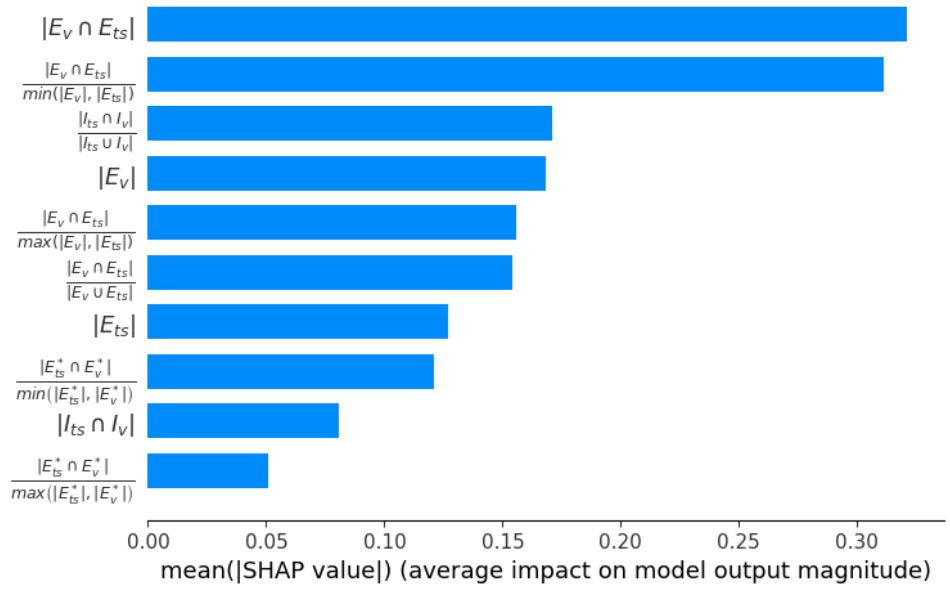

Fig. 2: Feature relevance analysis based on SHAP [46].

TABLE V: Retrieved video lists for query *Q1*

Matching videos ids	Ranked list (our approach)	Ranked list (JRHML)	Ranked list (JRL)	Ranked list (BERT-title+trans)
v1, *v2*, *v3*, *v4*, *v5*	**v3**	**v5**	**v5**	v6
	v1	**v3**	v7	v7
	v5	**v4**	**v2**	**v1**
	v2	v7	**v3**	v9
	v8	v9	v10	**v4**
	v4	**v2**	**v1**	**v3**
	v6	**v1**	v8	**v5**
	v9	v10	**v4**	v8
	v7	v6	v9	v10
	v10	v8	v6	**v2**
Average Precision	**0.81**	0.73	0.62	0.38

TABLE VI: Retrieved video lists for query *Q2*

Matching videos ids	Ranked list (our approach)	Ranked list (JRHML)	Ranked list (JRL)	Ranked list (BERT-title+trans)
v11, *v12*, *v13*	**v12**	v14	**v12**	**v11**
	v11	**v13**	v14	v15
	v16	v16	v16	v17
	v13	**v12**	v17	v18
	v15	**v11**	**v13**	**v12**
	v20	v21	**v11**	**v13**
	v14	v15	v19	v16
	v18	v18	v15	v14
	v21	v19	v21	v21
	v17	v17	v18	v19
Average Precision	**0.69**	0.28	0.60	0.21

TABLE VII: Retrieved video lists for query *Q3*

Matching videos ids	Ranked list (our approach)	Ranked list (JRHML)	Ranked list (JRL)	Ranked list (BERT-Title+Trans)
v21, *v22*, *v23*	**v21**	v24	v23	v25
	v22	v25	v24	v26
	v24	**v21**	**v21**	v24
	v23	**v23**	**v22**	v28
	v28	v27	v29	v30
	v27	v26	v30	**v22**
	v26	v28	v27	**v21**
	v25	v30	v28	**v23**
	v29	v29	v25	v27
	v30	**v22**	v26	v29
Average Precision	**0.69**	0.40	0.47	0.47

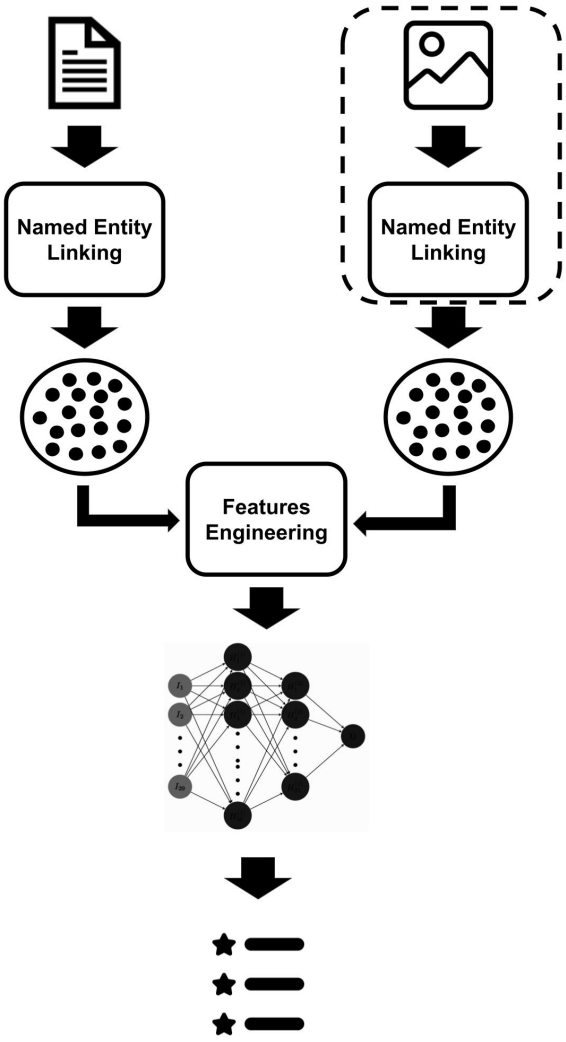

Fig. 3: Cross-media image retrieval task: from text to images.

The preliminary results show that our approach ranked third out of 14 tested methods. Thus, they confirm the portability of the proposed solution towards similar tasks and domains.

VI. Conclusions and Future Works

The paper presented a Deep Learning application aimed at supporting learners in the retrieval of educational videos pertinent to specific book fragments. To overcome the lack of metadata information in educational videos, it extracts NLP-based features from both the audio transcripts of the videos and the text snippets in the books. A Deep Neural Network model is trained on the textual descriptors of pairs of cross-media resources in order to produce the output rank pertinent to each input query.

The results indicate that the joint use of NLP and DL is appropriate while coping with educational resources with (partly) missing annotations. Furthermore, they show that the same approach can be successfully applied, with limited effort, to similar retrieval tasks and contexts.

In light of the research findings described above, as future work we plan to:

- Integrate the cross-media retrieval approach into interactive learning environments, such as mobile applications and serious games.
- Analyze the learners' user experience of the proposed system in various scenarios (e.g., higher level education, corporate training) through surveys, crowd-sourcing platforms, or interviews.
- Explore the portability of Deep Learning models to different contexts and media types.
- Design ad hoc Deep Learning architectures tailored to multilingual cross-media content.

References

[1] International Association of Universities, "Iau global survey on the impact of covid-19 on higher education around the world," 2020. [Online]. Available: https://www.iau-aiu.net/IMG/pdf/iau_covid-19_regional_perspectives_on_the_impact_of_covid-19_on_he_july_2020_.pdf

[2] UNESCO IESALC, "Covid-19 and higher education: today and tomorrow. impact analysis, policy responses and recommendations," 2020. [Online]. Available: http://www.guninetwork.org/files/covid-19_en_090420.pdf

[3] O. B. Adedoyin and E. Soykan, "Covid-19 pandemic and online learning: the challenges and opportunities," *Interactive Learning Environments*, pp. 1–13, 2020. [Online]. Available: https://doi.org/10.1080/10494820.2020.1813180

[4] R. Kay, "Exploring the use of video podcasts in education: A comprehensive review of the literature," *Computers in Human Behavior*, vol. 28, pp. 820–831, 05 2012.

[5] E. Bravo, B. García, P. Simo, M. Enache, and V. Fernandez, "Video as a new teaching tool to increase student motivation," 05 2011, pp. 638 – 642.

[6] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, 2018. [Online]. Available: https://doi.org/10.1109/TCSVT.2017.2705068

[7] D. Rao and B. McMahan, *Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning*. O'Reilly Media, 2019. [Online]. Available: https://books.google.it/books?id=3m69tAEACAAJ

[8] T. Duffy and D. Jonassen, *Constructivism and the Technology of Instruction: A Conversation*. Taylor & Francis, 2013. [Online]. Available: https://books.google.it/books?id=4QhGjtkobIUC

[9] P. Guo, J. Kim, and R. Rubin, "How video production affects student engagement: An empirical study of mooc videos," 03 2014, pp. 41–50.

[10] TedEd, "Ted ed lessons worth sharing." [Online]. Available: https://ed.ted.com/

[11] J. Kay, P. Reimann, E. Diebold, and B. Kummerfeld, "Moocs: So many learners, so much potential ..." *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 70–77, 2013.

[12] J. L. Martín Núñez, E. Tovar Caro, and J. R. Hilera González, "From higher education to open education: Challenges in the transformation of an online traditional course," *IEEE Transactions on Education*, vol. 60, no. 2, pp. 134–142, 2017.

[13] P. Kaur, H. S. Pannu, and A. K. Malhi, "Comparative analysis on cross-modal information retrieval: A review," *Computer Science Review*, vol. 39, p. 100336, 2021. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574013720304366

[14] H. Tong, J. He, M. Li, C. Zhang, and W.-Y. Ma, "Graph based multi-modality learning," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 862–871. [Online]. Available: https://doi.org/10.1145/1101149.1101337

TABLE VIII: Preliminary results on text-image retrieval. Wikipedia dataset.

Task	Our	BITR	CCA	CCASMN	CFA	CMCP	DCMIT	HSNN	JGRHML	JRL	LGCFL	ml-CCA	mv-CCA	S2UPG
Image→Text	0.331	0.222	0.249	0.246	0.277	0.326	0.277	0.321	0.329	0.339	0.274	0.269	0.271	0.377
Text→Image	0.251	0.171	0.196	0.195	0.226	0.251	0.250	0.251	0.256	0.250	0.224	0.211	0.209	0.286

[15] F. Wu, X. Lu, J. Song, S. Yan, Z. M. Zhang, Y. Rui, and Y. Zhuang, "Learning of multimodal representations with random walks on the click graph," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 630–642, 2016. [Online]. Available: https://doi.org/10.1109/TIP.2015.2507401

[16] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," in *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA*, M. desJardins and M. L. Littman, Eds. AAAI Press, 2013. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/view/6380

[17] G. Xu, X. Li, and Z. Zhang, "Semantic consistency cross-modal retrieval with semi-supervised graph regularization," *IEEE Access*, vol. 8, pp. 14 278–14 288, 2020.

[18] X. Zhai, Y. Peng, and J. Xiao, "Effective heterogeneous similarity measure with nearest neighbors for cross-media retrieval," in *Advances in Multimedia Modeling - 18th International Conference, MMM 2012, Klagenfurt, Austria, January 4-6, 2012. Proceedings*, ser. Lecture Notes in Computer Science, K. Schoeffmann, B. Mérialdo, A. G. Hauptmann, C. Ngo, Y. Andreopoulos, and C. Breiteneder, Eds., vol. 7131. Springer, 2012, pp. 312–322. [Online]. Available: https://doi.org/10.1007/978-3-642-27355-1_30

[19] D. Ma, X. Zhai, and Y. Peng, "Cross-media retrieval by cluster-based correlation analysis," in *IEEE International Conference on Image Processing, ICIP 2013, Melbourne, Australia, September 15-18, 2013*. IEEE, 2013, pp. 3986–3990. [Online]. Available: https://doi.org/10.1109/ICIP.2013.6738821

[20] Y. Zhang, Q. Ai, X. Chen, and W. B. Croft, "Joint representation learning for top-n recommendation with heterogeneous information sources," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, E. Lim, M. Winslett, M. Sanderson, A. W. Fu, J. Sun, J. S. Culpepper, E. Lo, J. C. Ho, D. Donato, R. Agrawal, Y. Zheng, C. Castillo, A. Sun, V. S. Tseng, and C. Li, Eds. ACM, 2017, pp. 1449–1458. [Online]. Available: https://doi.org/10.1145/3132847.3132892

[21] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang, "Cross-media semantic representation via bi-directional learning to rank," in *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*, A. Jaimes, N. Sebe, N. Boujemaa, D. Gatica-Perez, D. A. Shamma, M. Worring, and R. Zimmermann, Eds. ACM, 2013, pp. 877–886. [Online]. Available: https://doi.org/10.1145/2502081.2502097

[22] F. Wu, X. Jiang, X. Li, S. Tang, W. Lu, Z. Zhang, and Y. Zhuang, "Cross-modal learning to rank via latent joint representation," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1497–1509, 2015. [Online]. Available: https://doi.org/10.1109/TIP.2015.2403240

[23] J. Tang, Z. Li, M. Wang, and R. Zhao, "Neighborhood discriminant hashing for large-scale image retrieval," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2827–2840, 2015. [Online]. Available: https://doi.org/10.1109/TIP.2015.2421443

[24] F. Zhong, Z. Chen, and G. Min, "Deep discrete cross-modal hashing for cross-media retrieval," *Pattern Recognition*, vol. 83, pp. 64 – 77, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320318301924

[25] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 3441–3450. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298966

[26] S. Liu, S. Qian, Y. Guan, J. Zhan, and L. Ying, "Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, J. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, and Y. Liu, Eds. ACM, 2020, pp. 1379–1388. [Online]. Available: https://doi.org/10.1145/3397271.3401086

[27] S. ur Rehman, M. Waqas, S. Tu, A. Koubaa, O. ur Rehman, J. Ahmad, M. Hanif, and Z. Han, "Deep learning techniques for future intelligent cross-media retrieval," *CoRR*, vol. abs/2008.01191, 2020. [Online]. Available: https://arxiv.org/abs/2008.01191

[28] H. Bhuiyan, J. Ara, R. Bardhan, and M. R. Islam, "Retrieving youtube video by sentiment analysis on user comment," in *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 2017, pp. 474–478.

[29] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal transformer for video retrieval," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 214–229.

[30] F. Wang, C. Ngo, and T. Pong, "Structuring low-quality videotaped lectures for cross-reference browsing by video text analysis," *Pattern Recognit.*, vol. 41, no. 10, pp. 3257–3269, 2008. [Online]. Available: https://doi.org/10.1016/j.patcog.2008.03.024

[31] T. Tuna, J. Subhlok, L. Barker, V. Varghese, O. Johnson, and S. Shah, "Development and evaluation of indexed captioned searchable videos for STEM coursework," in *Proceedings of the 43rd ACM technical symposium on Computer science education, SIGCSE 2012, Raleigh, NC, USA, February 29 - March 3, 2012*, L. A. S. King, D. R. Musicant, T. Camp, and P. T. Tymann, Eds. ACM, 2012, pp. 129–134. [Online]. Available: https://doi.org/10.1145/2157136.2157177

[32] H. Yang and C. Meinel, "Content based lecture video retrieval using speech and video text information," *IEEE Transactions on Learning Technologies*, vol. 7, no. 2, pp. 142–154, 2014.

[33] A. Shoufan, "Estimating the cognitive value of youtube's educational videos: A learning analytics approach," *Computers in Human Behavior*, vol. 92, pp. 450 – 458, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0747563218301419

[34] G. Gimenez-Perez, N. Robert-Vila, M. Tomé-Guerreiro, I. Castells, and D. Mauricio, "Are youtube videos useful for patient self-education in type 2 diabetes?" *Health Informatics J.*, vol. 26, no. 1, 2020. [Online]. Available: https://doi.org/10.1177/1460458218813632

[35] A. E. Iannone, "Ballet education for the web 2.0 generation: A case for using youtube to teach elementary-school-aged ballet students," *Int. J. Technoethics*, vol. 10, no. 1, pp. 37–48, 2019. [Online]. Available: https://doi.org/10.4018/IJT.2019010104

[36] I. Radu, "Augmented reality in education: a meta-review and cross-media analysis," *Pers. Ubiquitous Comput.*, vol. 18, no. 6, pp. 1533–1543, 2014. [Online]. Available: https://doi.org/10.1007/s00779-013-0747-y

[37] S. Saurabh and A. S. Sairam, "Professors - the new youtube stars: education through web 2.0 and social network," *Int. J. Web Based Communities*, vol. 9, no. 2, pp. 212–232, 2013. [Online]. Available: https://doi.org/10.1504/IJWBC.2013.053245

[38] D. Vrandečić and M. Krötzsch, "Wikidata: A free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, p. 78–85, Sep. 2014. [Online]. Available: https://doi.org/10.1145/2629489

[39] M. Färber, F. Bartscherer, C. Menne, and A. Rettinger, "Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago," *Semantic Web*, vol. 9, pp. 1–53, 03 2017.

[40] A. Spitz, V. Dixit, L. Richter, M. Gertz, and J. Geiß, "State of the union: A data consumer's perspective on wikidata and its properties for the classification and resolution of entities," in *Wiki, Papers from the 2016 ICWSM Workshop, Cologne, Germany, May 17, 2016*, ser. AAAI Workshops, R. West, L. Zia, D. Taraborelli, and J. Leskovec, Eds., vol. WS-16-17. AAAI Press, 2016. [Online]. Available: http://aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13200

[41] M. Hadjieleftheriou and D. Srivastava, "Weighted set-based string similarity," *IEEE Data Eng. Bull.*, vol. 33, pp. 25–36, 2010.

[42] V. M K and K. K, "A survey on similarity measures in text mining," *Machine Learning and Applications: An International Journal*, vol. 3, pp. 19–28, 03 2016.

[43] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[44] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[45] M. Zhu, "Recall, precision and average precision," 09 2004.

[46] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[47] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *Transactions of Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 521–535, March 2014.