Knowledge-Driven Multimodal Named Entity Recognition in Audio and Video

***Abstract*—...**
***Index Terms*—...**

I. Introduction

Named Entity Recognition (NER) is an established task in the Natural Language Processing (NLP) and Knowledge Representation and Reasoning (KRR) fields. It uses textual and a priori knowledge to detect set of Named Entity (NE) and their types (e.g. a person). When extended to the Image Processing (IP), Audio Signal Processing (ASP) Computer Vision (CV) we talk about Multimodal Named Entity Recognition (MNER). Fig. 1 gives examples of Multimodal Named Entity (MNE) including different representations and types.

LOC, *PER* and *ORG* stand for location , person and organization

Fig. 1. Examples of MNEs

MNER has attracted an increasing attention in the research community over the last years. The primarily MNER methods process with text and image [16]. However, the Audio / Video (A/V) data are too popular means of information less considered in the MNER methods [1]. In this paper, we present a new MNER method in A/V having, as a particular property, to be knowledge-driven. The overall approach ensures a better robustness, scalability and optimization of the MNER task. Section 1 provides a state-of-the-art of MNER. Sections 2 and 3 detail our approach, experiments and results. Conclusions and perspectives are discussed in section 4. Table I gives the main symbols used in the paper.

II. State-of-the-art

III. The proposed approach

Fig. 2 gives the overview of our approach. From a Multimedia Knowledge Graph (MKG), we extract a set of Named Entity (NE) provided as text. These NEs are mapped to a database to identify candidate A/V data, called collections. A last, the detection of MNEs in A/V collections is performed in two steps, from coarse to fine. The next subsections detail the different components A. to E. of our approach.

TABLE I
Main symbols used in the paper

Symbols	Meaning
$\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l}, \mathbf{m}, \mathbf{n}$	integer variables
$\{\mathbf{NE_1}, \ldots, \mathbf{NE_m}\}$	a set of NEs
$\{\mathbf{h_1}, \ldots, \mathbf{h_n}\}$	a set of collection identifiers
$\mathbf{d_A}, \mathbf{d_V}$	durations of Audio and Video files
$\mathbf{\Delta}$	max duration for capture
$\mathbf{L} \in [\mathbf{L_{min}}, \mathbf{L_{max}}]$	the latency with $\mathbf{L_{min}} < \mathbf{0}$, $\mathbf{L_{max}} > \mathbf{0}$
$(\epsilon_1, \ldots, \epsilon_j)$	expert thresholds for A/V validation
$\mathbf{s}, \mathbf{e}$	time interval for capture with $\mathbf{s} < \mathbf{e}$ and $\mathbf{e} - \mathbf{s} = \mathbf{k} \times \mathbf{\Delta}$
$\mathbf{s^+}, \mathbf{e^-}$	corrected interval $\mathbf{s^+} = \mathbf{s} + \lvert\mathbf{L_{min}}\rvert$, $\mathbf{e^-} = \mathbf{e} - \mathbf{L_{max}}$
$\mathbf{t_0}, \mathbf{t1}$	start and end times of a broadcast event with $\mathbf{t_0} < \mathbf{t1}$
$\mathbf{t_0^-}, \mathbf{t1^+}$	corrected times $\mathbf{t_0^-} = \mathbf{t_0} - \lvert\mathbf{L_{min}}\rvert$, $\mathbf{t_1^+} = \mathbf{t_1} + \mathbf{L_{max}}$
$\mathbf{m} = \mathbf{f(n)}$	mathematic function for the collections and NEs
$\mathbf{T_L}, \mathbf{T_H}$	thresholds to filter the sets of NEs with $\mathbf{m} \in [\mathbf{T_L}, \mathbf{T_H}]$

I provide all, we will filter the baseline later

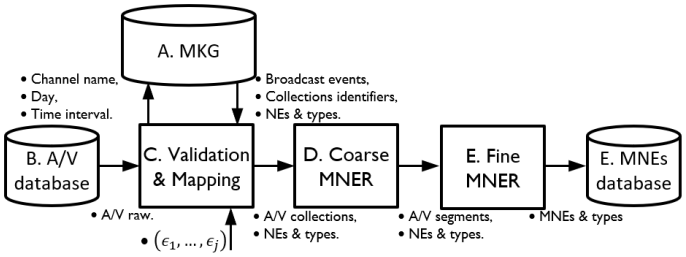

Fig. 2. The proposed approach

A. The multimedia knowledge graph

As discussed in sections 1 and 2, we propose in this paper a knowledge-driven MNER in A/V data. This ensures a better robustness, scalability and optimization of the MNER task. Multimedia knowledge representation is a well known topic in thE literature. Different approaches can be used for the formalization. The ontology is the most common typically represented as a Multimedia Knowledge Graph (MKG) [5]. A MKG can represent a large variabilities of entities and types. In our work, we have used the public MKG detailed in [21]. This MKG has a particular focuss on the NEs representation. That is, it is fully dedicated to our MNER task.

TABLE II
The Multimedia Knowledge Graph (MKG) [21]

Start	Length	Broadcast Events	Collections	NEs
10/2023	1 year	2.9M	70K	84K

The KG must be extended to 04/2025 for the full dataset

This MKG has been designed from Francophonie Electronic Program Guide (EPG) data of Television (TV). An overall

pipeline has been designed for the Named Entity Recognition (NER) and Linking (NEL) from text information contained in EPG data. Fig 3 illustrates the organization of data within the MKG. The data are provided as TV collections having an unique identifier $\mathbf{h}$. Each collection contains a set of TV Broadcast Event (BCE) (e.g. episodes of a same show). A collection could have a large amount of BCEs and then a long total duration. Every collection is linked to a specific set of NEs $\{\mathbf{NE_1}, \ldots, \mathbf{NE_m}\}$. As detailed in Table II, the MKG contains 84K unique NEs dispatched into 70K collections.

Collection:	TVC0001578	TVC0007731
Title:	Des trains pas comme ..	Grands Reportages
Channel:	France 5	TFI
Category:	Cultural	Politics
BCEs:	945	85
Duration:	850h	105h
NEs:	230	20

Fig. 3. Organization of data in the knowledge database [21]

B. The A/V database

Our knowledge database is dedicated to French TV content. A corresponding database, containing French TV A/V data, must captured for mapping. A first solution is to request catch-up services [22]. This has a restricted content (e.g. popular shows and channels) and a low quality of A/V data[1]. The alternative is the live broadcasting ensuring a scalable and continue capture at a high quality. This requires a workstation, in our work we have used the one detailed in [17]. For the needs of clarification of our work, we introduce it for short.

In the workstation Fig. 4, the A/V data is obtained from the French Digital Terrestrial Television (DTT) signal processed with a multituner (i) and delivered to a computer. With a DTT capture, the A/V data has no latency compared to catch-up services on network. The video is re-encoded in real-time with dual-channel cards (ii), whereas the audio is encoded by CPU (iii). The computer embeds 4 cards able to capture 8 channels at a time. It is set with a high processing and memory capacity (iii) (iv) for A/V processing and storage.

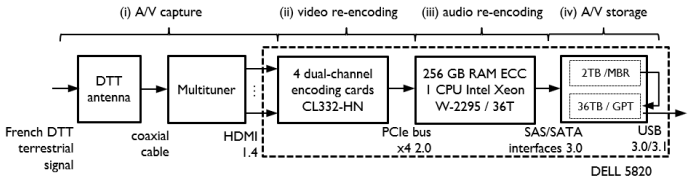

Fig. 4. The TV Workstation [17]

Table III gives our capture setting derived from [17]. For the needs of a good visual rendering and memory cost optimization, we encode the video at the Standard Definition (SD) resolution with a high Frames Per Second (FPS) and Mbps. The audio is encoded at the highest quality supported by the French DTT. The capture is triggered daily on the interval $\mathbf{s}$ =4am (start) to $\mathbf{e}$ =12pm (end) considering the night interruption of TV broadcasters. The workstation captures daily $160 = 20 \times 8$ hours and $\simeq 128$ GB of A/V data.

[1]streaming latency, heterogeneous A/V encoding from catch-up services.

TABLE III
THE CAPTURE SETTING

Video			Audio	Daily ($\mathbf{s}$ =4am to $\mathbf{e}$ =12pm)			
Resolution	Mbps	FPS	Kbps	A/V	GB	Δ	Files
720×576	1.6	30	256	160h	127.9	5h	32×2

As detailed in Fig. 4, the A/V encoding is two-steps in the workstation. The video is encoded in hard real-time with the capture cards (ii), whereas the audio is encoded in soft with the CPU (iii). The audio encoding suffers from latency that could be expressed as $\mathbf{L} = \mathbf{d_A} - \mathbf{d_V}$ (with $\mathbf{d_A}, \mathbf{d_V}$ the durations of audio and video files). This results in most of the time in $\mathbf{L} < \mathbf{0}$. If this latency has none impact for a separate analysis of A/V data [17], it raises a problem for a multimodal processing. To deal with it, we reset the capture with after a duration $\mathbf{\Delta} = 5$ hours (with $\mathbf{e} - \mathbf{s} = \mathbf{k} \times \mathbf{\Delta}$). This ensures a good tradeoff between the latency maximization and fragmentation of A/V files. 32×2 A/V files are then captured daily by the workstation.

Table IV details our A/V database. We have set the workstation to capture 8 main French TV channels, for a near 2 months, on the period February to April 2025. The database contains a near 10 thousands hours and 7.8 TB of A/V data.

TABLE IV
THE A/V DATABASE

Start	Length	Duration	Files	TB
Feb. 2025	9 weeks	10,080h	2016×2	7.8

C. The validation and mapping

From the knowledge and A/V databases, the next step of our approach is mapping Fig. 2. The goal is to identify the relevant A/V data to be associated to the right sets of NEs. The mapping is done in two steps as detailed below.

Validation of A/V data: a huge amount of A/V files has been captured with the workstation Table IV. These could suffer from encoding troubles. To ensure the quality, we control the properties of A/V files (e.g. $\mathbf{d_A}, \mathbf{d_V}$, Mbps, Kbps, FPS) with $\min / \max$ thresholds $(\epsilon_1, \ldots, \epsilon_j)$. All the files out of the ranges are deleted from the A/V database. In addition, some A/V files could suffer of latency $\mathbf{L}$. From the latency distribution, we establish $\mathbf{L}_{\min}, \mathbf{L}_{\max}$ to filter out these files.

Mapping with the timing information: the captured A/V data is given with information (days, start/end times $\mathbf{s}, \mathbf{e}$ of the capture and channel names). From this, a query is submitted to the KG for every daily A/V capture of a channel (i.e., $\simeq 500$ queries in total). Every query is established to return a set of consistent BCE between the KG and A/V data. The consistency guarantees the presence of BCEs in A/V data (when $\mathbf{t_0} \geq \mathbf{s}$ and $\mathbf{t_1} \leq \mathbf{e}$, with $\mathbf{t_0}$, $\mathbf{t_1}$ the timestamps of a

BCE). However, the TV broadcasting suffers from latency $\mathbf{L}$ that can be modelled as a gaussian distribution [17] (with $\mathbf{L} \in [\mathbf{L}_{\min}, \mathbf{L}_{\max}]$ and $\mathbf{L}_{\min} < \mathbf{0}, \mathbf{L}_{\max} > \mathbf{0}$). For a full consistency, $\mathbf{t_0}$, $\mathbf{t_1}$ must be redefined as $\mathbf{t_0^-} = \mathbf{t_0} - |\mathbf{L}_{\min}|$ and $\mathbf{t_1^+} = \mathbf{t_1} + \mathbf{L}_{\max}$. To embed the latency parameters in queries, we redefine $\mathbf{s}$, $\mathbf{e}$ as $\mathbf{s}^+ = \mathbf{s} + |\mathbf{L}_{\min}|$, $\mathbf{e}^- = \mathbf{e} - \mathbf{L}_{\max}$.

Mapping with the collection identifiers and NEs: every concistent BCE returned in step one is provided with a unique collection identifier $\mathbf{h}$. We aggregate these identifiers only once to constitute the set $\{\mathbf{h_1}, \ldots, \mathbf{h_n}\}$. This set is submitted as a query to the KG. We obtain for every collection $\mathbf{h}$ its corresponding set of NEs $\{\mathbf{NE_1}, \ldots, \mathbf{NE_m}\}$. The distribution of NEs among the collections could be formalized as a function $\mathbf{m} = \mathbf{f}(\mathbf{n})$. Two extrema appear when $\mathbf{m} \to \mathbf{0}$ and $\mathbf{m} \to +\infty$. The first corresponds to the collections having none or too few NEs. This is out of interest for the MNER task and results in processing overhead. The second is obtained for the collections having a high scalability. Here, the knowledge-driven MNER is out of support. To deal with this problem, we process with thresholding to select the relevant collections having $\mathbf{m} \in [\mathbf{T_L}, \mathbf{T_H}]$. The $\mathbf{T_L}, \mathbf{T_H}$ thresholds are obtained from the mathematic analysis of the $\mathbf{f}(\mathbf{n})$ interpolated function.

Next to the thresholding, we obtain a new set of collections $\{\mathbf{h_1}, \ldots, \mathbf{h_{\tilde{n}}}\}$ (with $\tilde{\mathbf{n}} \ll \mathbf{n}$). All the consistent BCEs, linked to these collections, are used to map the A/V data. The captured A/V files (having a duration $\mathbf{\Delta} = \mathbf{5}$ hours) are segmented from timestamps of BCEs $\mathbf{t_0^-}, \mathbf{t_1^+}$ (having a duration $\mathbf{d_V} = \mathbf{d_A} = \mathbf{t_1^+} - \mathbf{t_0^-}$). A merging process is applied for any BCE appearing between two consecutive captured files. The obtained files are dispatched among different directories corresponding to the collections where each directory / collection is linked to its set of NEs $\{\mathbf{NE_1}, \ldots, \mathbf{NE_m}\}$.

D. The coarse MNER

E. The fine MNER and database

IV. Experiments and results

We present first the results of the validation step. As detailed in section III-C, this is done with expert setting using $\min / \max$ thresholds $(\epsilon_\mathbf{1}, \ldots, \epsilon_\mathbf{j})$ to be applied to the properties of A/V files. They are detailed in Table V.

TABLE V
Validation of A/V database

$(\epsilon_\mathbf{1}, \ldots, \epsilon_\mathbf{j})$	$\mathbf{d_V}$	**Mbps**	**FPS**	$\mathbf{d_A}$	**Kbps**	**L**
min	18014s	1.6		17996s	255.5	-18.07
max	18020s	2.1		18002s	255.55	-17.54
η	0.03%	NA		0.03%	0.02%	3.03%

The video files are captured using the parameter $\mathbf{\Delta}$ = 5h = 18,000s. The cards in the workstation (section III-B) encode with an overflow due to the trigger and closure processes. The video durations enter mainly in the range $\mathbf{d_V} \in [18014, 18020]$s having a low variation gap of $\eta = 0.03\%$ (6s at worst). All the files out of that range result of capture troubles not validated in the A/V database. The cards are set with an encoding parameter of 1.6 Mbps (Table III). The real-time encoding is adaptive that could result in higher Mbps $\in]1.6, 2.1]$. The highest quality videos are out of impact for the MNER task. However, all the videos with a lowest encoding < 1.6 Mbps are unvalidated in the A/V database.

The encoding process for the audio files enters in a same template. The audio duration has a main range $\mathbf{d_A} \in [17996, 18002]$s with a same variation gap of $\eta = 0.03\%$ (6s at worst). The files out of that range are unvalidated. The audio files are CPU encoded without adaptation in the workstation (see section III-B). They have a near constant Kbps $\simeq$ 255.5 (with a $\eta = 0.02\%$). The files with a lowest Kbps are unvalidated in the A/V database. The main distribution of the latency $\mathbf{L} = \mathbf{d_A} - \mathbf{d_V}$ is fully negative in the range $\mathbf{L} \in [-18.07, -17.54]$s. It has a variation of $\eta \simeq 3\%$ corresponding to a $\frac{1}{2}$s gap at worst between A/V files.

In a final step, using all the using $\min / \max$ thresholds $(\epsilon_\mathbf{1}, \ldots, \epsilon_\mathbf{j})$ Table V, any couple of A/V files unvalidated (audio AND/OR video) is deleted from the A/V database.

V. Conclusions and perspectives

APPENDIX A

TABLE VI
CATEGORIZATION OF THE MULTI-MODALITY AND SCALABILITY

Ref	Year	Lang	A	V	T	E	Cat	Length	NEs		Ty
									Total	U	
[14]	2005	English	✓	✓		✓	1	$\gg$ 620h		93	3
[4]	2012	Dutch				✓	$\geq$1	$\ll$10Kh			4
[11]	2013	English		✓	✓	✓	9	135h	44.5K		10
[6]	2014	English			✓	✓	1				10
[7]	2014	English		✓	✓	✓	1	$\simeq$ 350h			10
[20]	2014	English	✓	✓	✓	✓	1	2h		69	12
[10]	2015	French		✓		✓	1		2K	285	1
[15]	2021	Chinese		✓			1	160h		66	1
[3]	2021	English		✓	✓		1	$\simeq$ 1h			1
[1]	2023	English		✓	✓		1	75h	126K	15	1
[18]	2024	English		✓	✓		1	367h	197K	??	??
[2]	2024	English		✓		✓	??	??	??	??	??
[19]	2025	English	✓	✓	✓		13	1h	0.2K	??	45
	20xx	??					??	??	??	??	??

Ref, Lang, Cat, U and Ty stand for Reference, Language, Category, Unique and Types
A/V/T/E stand for Audio, Video, Text and External

TABLE VII
CATEGORIZATION OF APPROACHES

Task	Approaches
Extraction of text entities	• Speech-To-Text [14], [20] • Video OCR [3], [14]
Text processing	• Textual metadata [4] • Web sources [6], [10], [14], [20] • Named Entity Recognition [1], [3], [4], [6], [7], [11], [14], [20] • Resolution [6], [14] • Validation with ontology [6], [7], [11]
Extraction of visual entities	• Face detection & recognition [1], [15], [20] • Logo recognition [10]
Connecting entities	• Connecting [14] • Temporal alignment [7], [11], [15], [20]

TABLE VIII
CATEGORIZATION OF APPROACHES

Audio Signal Processing	Knowledge Representation
• Speech-To-Text [14], [20]	• Web sources [6], [10], [14], [20] • Textual metadata [4] • Ontology [6], [7], [11]
Natural Language Processing	**Computer Vision**
• Statistical model [14] • Text features [4], [20] • Classification [4] • Framework / library [1], [3], [6], [7], [11] • Resolution [6], [14] • Statistical analysis [6], [11], [20] • Connecting [14]	• OCR [3], [14] • Face detection & recognition [1], [15], [20] • Logo recognition [10]

• PapernickN2005 [14]: NER in broadcast news, the approach processes with transcripts (A) "Dataset Informedia Project web site at Carnegie Mellon University, http://www.informedia.cs.cmu.edu" for training and video OCR (V), the paper is pure NLP. Every query is returned as a link graph of relevant entities, where the user can manipulate and browse (with access to the NEs and video clips). The NLP method processes with statistical language modeling, a Viterbi algorithm then finds the best path through the named-entity options. A core algorithm is used to map the different forms "Entity Resolution" (e.g. George Bush, George W. Bush and Bush) into a common representation. An approach is proposed to understand the relational context between NEs, relational context are then expressed into a graph. Experiments testing are done on 622 hours of CNN brodcast news from 93 unique NEs (extracted from an external source infoplease.com's list). . 3 types are used (PER, LOC, ORG), the information about the amount of detected NEs in the 622 hours is unclear.

• DeleuJ2012 [4]: NER on multimedia archives in Dutch (Flemish). The considered data consists of archivation metadata from video collections (VLIB is the Flemish research project providing the dataset, it contains roughly about 10,000 hours of video material with a large amount of textual metadata). 4 types are used (PER, LOC, ORG and MISC). The NER is done on the metadata part (none on the audio and video), it uses features (word, shape, character n-grams and conjunctions) and classification (Conditional Random Fields "CRFs"). Words and phrase clustering are used to make robust the NER and NLP methods to new vocabulary. Specific process for capitalization and word/phrase clusters are proposed. The experiments are done 1000 documents of the VLIB dataset (the total number of document in the dataset is unknown). None information is given about the UNEs and NEs, but as only the metadata is processed, the scalability looks then very small.

• LiY2013 [11]: it uses textual and video features (named entities extracted from subtitles, temporal features / duration of the media fragments where entities are spotted). The dataset is composed of 805 Dailymotion video (web sources) having a duration of 135h (484417s). The video collection is organized among the dailymotion channels into 9 categories (fun, tech, sport, news, creat, lifes, films, music, other). The NER in subtitle is multilingual, based on the NERD framework[2]. The entities are classified using the core NERD Ontology / LinkedTV Ontology, there is then an external web source used here. The NEs are classified into 10 types (Thing, Amo, Ani, Evt, Func, Loc, Org, Person, Prod, Time). A total of 44.5K NEs are tested (= 8961 + 7049 + 178 + 124 + 1100 + 4262 + 4030 + 11976 + 3595 + 3201). The NEs are aligned with the video segment with a grouping algorithm. Into the deep, the paper details the multiclass classification problem (with machine learning algorithms are used for the multiclass classification "Logistic Regression (LG), K-Nearest Neighbour (KNN), Naive Bayes (NB) and Support Vector Machine (SVM)") to address 4 main statistical questions baout the correlation of NEs / Types with the videos ... a good and complete paper.

• GarciaJLR2014a [6] "Augmenting TV Newscasts ...": the system collects information about persons, locations, organizations and concepts (4 types) occurring in the newscast videos. It processes with NER on subtitles, based on the NERD framework[3]. The detected NEs are liking using web identifiers (i.e. for Entity Disambiguation) using API (AlchemyAPI, DBpedia, Spotlight). The NEs are notated using the NERD Ontology and ranked for selection (most popular). NEs not only occur in the subtitles but can also be related to the video indirectly. For solving, NER is performed on additional Web

[2]NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools, 2013

[3]NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools, 2013

documents related to the main video. A method for querying web document is proposed for linking. The NER process is applied to the new documents with clustering and reranking of NEs. There is no process of videos, just a mapping of detected NEs to them. No information is given about the video categories and experiments. The paper is linked to the LinkedTV EU project[4].

• GarciaJLR2014b [7] "Detecting Hot Spots ..." : This paper processes with educational online videos. The video data are processed with a shot algorithm (temporal feature). Next, the approach works from subtitles. The dataset is composed of 1681 TED talks (https://www.ted.com/) $\simeq$ 12 minutes per talk, $\simeq$ 350h. The NER in subtitle is multilingual, based on the NERD framework[5]. Similar timing segments (topics, NEs) are merged with a similarity function. No detail are given about the UNEs and NEs. The paper is linked to the LinkedTV EU project[6].

• ShresthaN2014 [20]: the works is dedicated to British royal wedding videos using automatic speech recognition (ASR), subtitle and visual data. An external source or articles (9 Wikipedia article) is used to get external NEs. From articles, NEs are obtained with linguistic features and data mining technique to find frequent word patterns. Results are compared on transcripts vs. subtitle (WinDiff, P/R/F scores). Videos are processed with face detection (Bradski, 2000) and matching. A bag-of-visual-words features are extracted from detected faces, clustering is applied. For any key-frame containing a face, NEs appearing in common in transcript (get with ASR) and subtitles. For NEs / face alignement, EM algorithm is applied. Experiments is done on a single video of 2 hours, with UNEs of 69 but no information about NE in A/V data. The types are two, mainly people and events.

• LetessierP2015 [10] system to detect NEs / logos in images and videos, among a list of 25K entities, aggregated from Wikipedia lists / specific websites. The system is dedicated to logo indentification. It process with scene text / logo detection with SIFT and classification (based on matching function). The system was tested on a database of 2K images / containing 285 NEs. there is no real experiments on videos, mainly images, video was just for testing.

• QianH2021 [15]: the work aims at scene information detection method based on entity recognition, using spatiotemporal relationships. There is 4 subcategories of drama TV series (medical, serie, fantasy, romance). Only the video data is processed here. The Experiments are done on TV series data. The video objects are obtained with face detection and recognition, then labelled as subjects if well recognized (or objects if not). Yolo V4 is used for face detection, then a Face recognition algorithm is applied. The key contribution of the paper is a scene detection algorithm processing with a Bayesian probability model (Entity relationship algorithm). It processes in 3 stages (Spatiotemporal Relationship object/subject, Abnormal case recognition, Scene Detection). 5 datasets are used for experiments for the recognition of $66 = 7 + 10 + 18 + 17 + 14$ NEs / face, that is Type is 1 for PER. No information is given about the dataset size, external check said 160h.

• CaoY2021 [3]: the works is dedicated to educational videos. It propose a complete framework for active Viewing and visual note-taking, where the goal is to produce rendering stickers summarizing the videos. A part of the paper is dedicated to MNEs. The educational video contains many figures, tables, equations, flowcharts, etc. The systems processes videos to extract automatically text with OCR (Google's Cloud Vision API). The work processed too with manuel transcript (transcript of the speech with additional information, it looks an external metadata about the video). Some visual notes (output of user interaction for frame and object sticker selection) are as well produced by users. The NLP library Spacy is applied to the transcripts to get NEs. 11 videos are used for testing (no information about NEs, UNEs and Ty) for a $\simeq$ 1h (7.5 + 8.5 + 8.5 + 2.75 + 4.5 + 5.5 + 8 + 14.5 + 3.5 + 3.5 + 21.5). It looks to have no correlation between the OCR, transcript and user note for NE analysis.

• AyoughiM2023 [1]: it is dedicated to visual named entity discovery in videos, considering 2 types (face, object). The system processes in 3 steps (i) bipartite entity-name graphs from frame-caption pairs (ii) detecting visual entity agreements (iii) refine the entity assignment. The step (i) is done NER tool (a BERT fine tuned for English) on subtitles then video, it results in key-frames (how text NEs and key-frame alignement?), with a baseline face detection (github project). It produces pair of text NEs + RoI / face. The step (ii) processes with clustering to aggregate the RoI / face. The step (iii) deals with the case of unbalance distribution of visual NEs to find the most significant cluster centroid. Experiements are done from subdataset of TV-QA (with a hight quality manual transcripts, including text, timestamps). Experiments are given on SC-friends (5 episode of season 1 friends) and SC-BBT (all episodes of Big Bang Theory seasons 1 to 9, plus the 13 first episode of season 13). That is, the two dataset cover a near 75h of A/V data (4500 seconds). The SC-Friends cover 3.3K (3386) visual NEs for 7 UNEs (Joey, Rachel, Ross, Phoebe, Monica, Chandler, and an unknown), the SC-BBT 122.2K (122264) for 8 main characters or unknown.

[4]https://mklab.iti.gr/projects/linkedtv/

[5]NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools, 2013

[6]https://mklab.iti.gr/projects/linkedtv/

REFERENCES

[1] M. Ayoughi and al. Self-Contained Entity Discovery from Captioned Videos. ACM Transactions on Multimedia Computing Communications and Applications, vol. 19(5), No 177, 2023.

[2] H.A. Ayyubi and al. VIEWS: Entity-Aware News Video Captioning. Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 20220–20239, 2024.

[3] Y. Cao and al. VideoSticker: A Tool for Active Viewing and Visual Note-taking from Videos. International Conference on Intelligent User Interfaces (IUI), pp. 672–690, 2022.

[4] J. Deleu and al. Named Entity Recognition on Flemish audio-visual and newspaper archives. Workshop on Dutch-Belgian Information Retrieval (DIR), p.38-41, 2012.

[5] H. Fang and al. Cross Attention Scoring Function for Multimedia Domain Knowledge Graph Completion. Symposium on Broadband Multimedia Systems and Broadcasting (BMSB),pp. 1-6, 2023.

[6] J.L.R. Garcia and al. Augmenting TV Newscasts via Entity Expansion. In The Semantic Web (ESWC), pp. 472–476, 2014.

[7] J.L.R. Garcia and al. Detecting Hot Spots in Web Videos. International Semantic Web Conference (ISWC), Vol. 1272, 2014.

[8] S. Ghannay and al. End-to-end named entity extraction from speech. Spoken Language Technology Workshop (SLT), pp. 692-699, 2018.

[9] X. Hao and al. Intro and Recap Detection for Movies and TV Series. Winter Conference on Applications of Computer Vision (WACV), pp. 167-176, 2021.

[10] P. Letessier and al. DigInPix: Visual Named-Entities Identification in Images and Videos. International Conference on Multimedia Retrieval, pp.661-664, 2015.

[11] Y. Li and al. Enriching Media Fragments with Named Entities for Video Classification. International Conference on World Wide Web (WWW), pp. 469–476, 2013.

[12] S. Mdhaffar and al. End-to-end model for named entity recognition from speech without paired training data. Interspeech conference, 2022.

[13] Q. Meeus and al. MSNER: A Multilingual Speech Dataset for Named Entity Recognition. Workshop on Interoperable Semantic Annotation (ISA), pp. 8-16, 2024.

[14] N. Papernick and al. Summarization of Broadcast News Video through Link Analysis of Named Entities. The AAAI Workshop on Link Analysis, 2005.

[15] H. Qian and al. Video Scene Information Detection Based on Entity Recognition. Wireless Communications and Mobile Computing, Vol. 2021, Num 1020044, 2021.

[16] S. Qian and al. A Survey on Multimodal Named Entity Recognition. International Conference on Advanced Intelligent Computing Technology and Applications (ICIC), pp. 609-622, 2023.

[17] F. Rayar, M. Delalandre and V.H. Le. A large-scale TV video and metadata database for French political content analysis and fact-checking. Conference on Content-Based Multimedia Indexing (CBMI), pp. 181-185, 2022.

[18] K. Dela Rosa. Structured Entity Extraction from Travel Videos Using Vision-Language Models. Workshop on Recommenders in Tourism (RecTour), pp. 23-30, 2024.

[19] K. Dela Rosa. RAVEN: An Agentic Framework for Multimodal Entity Discovery from Large-Scale Video Collections. In Open-access repository (arXiv), No 2504.06272, 2025.

[20] N. Shrestha and al. Key Event Detection in Video using ASR and Visual Data. Workshop on Vision and Language (VL), 2014.

[21] H.G. Vu, N. Friburger, A. Soulet and M. Delalandre. stvd-kg: A Knowledge Graph for French Electronical Program Guides. International Semantic Web Conference (WISE), 2025.

[22] Y. Xu and al. Watching the Watchers: Automatically Inferring TV Content From Outdoor Light Effusions. Conference on Computer and Communications Security (CCS), pp 418–428, 2014.