Creating Enriched YouTube Media Fragments With NERD Using Timed-Text

Yunjia Li[1], Giuseppe Rizzo[2], Raphaël Troncy[2], Mike Wald[1], and Gary Wills[1]

[1] University of Southampton, UK
yl2@ecs.soton.ac.uk, mw@ecs.soton.ac.uk, gbw@ecs.soton.ac.uk
[2] EURECOM, Sophia Antipolis, France,
giuseppe.rizzo@eurecom.fr, raphael.troncy@eurecom.fr

Abstract. This demo enables the automatic creation of semantically annotated YouTube media fragments. A video is first ingested in the Synote system and a new method enables to retrieve its associated subtitles or closed captions. Next, NERD is used to extract named entities from the transcripts which are then temporally aligned with the video. The entities are disambiguated in the LOD clound and a user interface enables to browse through the entities detected in a video or get more information. We evaluated our application with 60 videos from 3 YouTube channels.

Keywords: Media fragment, media annotation, NER

1 Introduction

New W3C standards such as HTML5, Media Fragment URI and the Ontology for Media Resources have finally made videos a first class citizen on the Web. Indexing a video at a fine grained level such as the scene is, however, not yet a common practice on popular video sharing platform. In this demo, we propose to use NERD for extracting named entities from timed text associated to videos in order to generate media fragments annotated with resources from the LOD cloud. Our contributions include a new combined strategy for extracting named entities, temporal alignment of the named entities with the video and a user interface for browsing the enriched videos.

The LEMO multimedia annotation framework provides a unified model to annotate media fragments while the annotations are enriched with contextually relevant information from the LOD cloud [1]. Yovisto provides both automatic video annotations based on video analysis and collaborative user-generated annotations which are further linked to entities in the LOD cloud with the objective to improve the searchability of videos [5]. SemWebVid automatically generates RDF video descriptions using their closed captions [4]. The captions are analyzed by 3 web services (AlchemyAPI, OpenCalais and Zemanta) but chunked into blocks which make loose the context for the NLP tools. In this demo, we propose a new combined strategy using 10 different NER tools based on NERD [3]. In addition, we propose a new method to get the subtitles of a video and to analyze them globally while re-creating the temporal alignment.

2 Technical Architecture

This demo is powered by the integration and extension of two systems: Synote [2] and NERD [3] (Figure 1a). A user creates a new recording in Synote from any

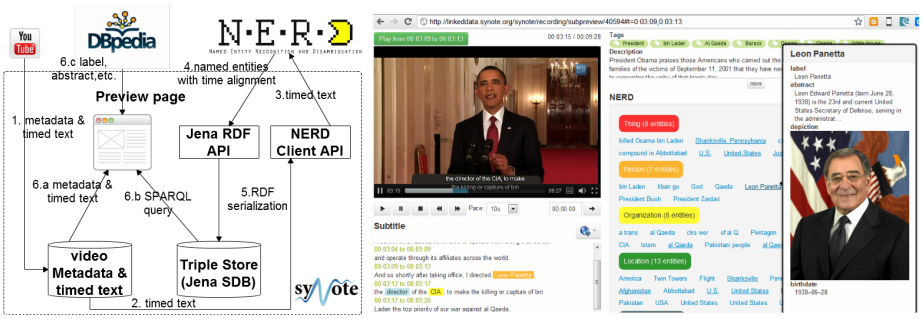

Fig. 1. a) Synote and NERD integration architecture. b) The Synote UI enriched with NERD and DBpedia

YouTube video. **(1)** The system first extracts the metadata and the subtitles if available using the YouTube API:

```
GET api/timedtext?v=videoid&lang=en&format=srt&name=trackname
```

In this request, four parameters are required: the YouTube video id *v*, the language of the subtitles *lang*, the timed-text format *format* and the track *name*. **(2)** A prior request is necessary for getting the track name since it is specified by the video owner.

```
GET api/timedtext?v=videoid&type=list
```

(3) The timed text is passed to the NERD client API which sends it to the NERD server. The named entity extraction is then performed on the entire context of the SRT file. **4** NERD returns a list of named entities with their type and a URI that disambiguates them, and a temporal window reference *startNPT* and *endNPT* corresponding to the SRT block where the entity appears. NERD exploits a combined strategy where 10 different extractors are used together. The named entity types are aligned yielding to a classification in 8 main types plus the general *Thing* concept. **5** On receiving the NERD response, Synote constructs media fragment URIs and uses the Jena RDF API to serialize the fragment annotations in RDF. The vocabularies NERD[3], Ontology for Media Resource[4], Open Annotation[5] and String Ontology in NIF[6] are used. Finally, the user interface shows the linking between named entities and media fragments, together with the YouTube video and interactive subtitles. The named entities and related metadata extracted from the subtitles are retrieved through

[3] http://nerd.eurecom.fr/ontology
[4] http://www.w3.org/ns/ma-ont
[5] http://www.openannotation.org/spec/core
[6] http://nlp2rdf.lod2.eu/schema/string

SPARQL queries **(6.a, 6.b)**. If a named entity has been disambiguated with a dbpedia URI **(6.c)**, a SPARQL query is sent to get further data about the entity (e.g. label, abstract, depiction) which is displayed alongside with the named entities.

3 Walk Through Demo

A live demo can be found at http://linkeddata.synote.org[7]. A user first logged in on Synote. When going to the recording creation page, a user can start the ingestion of a YouTube video. The recording is then available in the recording list. The "NERD Subtitle" button enables to launch the named extraction process. When completed, a "Preview Named Entities" button enables to go to the player page where named entities can be used to seek in particular video fragments.

Figure 1b shows the screenshot of a preview page. The right column displays the named entities found grouped according to the 8 main NERD categories. The YouTube video is included in the left column together with the interactive subtitles. The named entities are highlighted in different colours according to their categories. If a media fragment is used in the preview page URI, the video starts playing from the media fragment start time and stops playing when the end time is reached. When clicking on an named entity, the video jumps to the media fragment that corresponds to the subtitle block where the named entity has been extracted. If a named entity has been disambiguated with a dbpedia URI, the entity is underlined. In addition, when the entity is hover, a pop-up window shows additional information such as the generic label, abstract and depiction properties. For named entities of type Person, the birth date is displayed while latitude and longitude information are given for Location.

4 Evaluation

We filtered the videos which have subtitles for 3 different channels: *People and Blogs*, *Sports* and *Science and Technology* and collected 60 videos in total (the top 20 for each category). Videos have different duration ranging from from 32 to 4505 seconds and different popularity ranging from 18 to 2,836,535 views (on July 30th, 2012). The corpora is available at http://goo.gl/YhchP and can be visually explored in Synote at http://goo.gl/XmMqp after being logged in with the iswc2012 account. The video #16 is the only one discarded because its subtitles are written in Romanian. The evaluation consisted in two steps: *i)* be able to get all subtitles and *ii)* perform entity recognition using NERD. We combined all extractors supported by NERD and we aligned the classification results to 8 main types (*Event* is only supported by OpenCalais in beta) plus the general type *Thing* used as fallback in the case NERD cannot find a more specific type. We define the following variables: number of documents per category n_d; total number of words n_w; number of words per document ratio r_w; total number of entities n_e; number of entities per document r_e (Table 1). We observe that *Science and Technology* videos tend to be more about people and organizations while *Sports* videos mention more often locations, time and amount. *People and Blogs* videos have less useful information although it is interesting to see that this type of video can be used to train event detection.

[7] As credentials, please insert for both user and password: "iswc2012".

	People and Blogs	Sports	Science and Technology
n_d	19	20	20
n_w	7,187	21,944	39,661
r_w	378.26	1,097.20	1,983.05
n_e	610	897	1,303
r_e	32.11	44.85	65.15
Thing	6.68	15.35	14.75
Person	4.42	9.75	**14.55**
Function	0.74	**7.35**	1.15
Organization	3.63	9.20	**12.25**
Location	3.89	**8.05**	6.40
Product	3.26	2.60	**6.40**
Time	3.95	**13.80**	3.35
Amount	5.47	**9.30**	6.30
Event	**0.05**	0.00	0.00

Table 1. Upper part shows the average number of named entities extracted. Lower part shows the average number of entities for the 8 NERD top categories grouped by video channels.

5 Conclusion

This paper presents a demo that creates media fragments from YouTube videos and annotates their subtitles using NERD. The process extracts entities which are contextualized for the entire Timed-Text document and consequently enriches media fragments with pointers to the LOD cloud. Finally, we provide an informative analysis of the type of entities one can expect depending on video channels which will require much more thorough investigation.

References

1. Haslhofer, B., Jochum, W., King, R., Sadilek, C., Schellner, K.: The LEMO annotation framework: weaving multimedia annotations with the web. International Journal on Digital Libraries 10(1), 15–32 (2009)
2. Li, Y., Wald, M., Omitola, T., Shadbolt, N., Wills, G.: Synote: Weaving Media Fragments and Linked Data. In: 5^{th} International Workshop on Linked Data on the Web (LDOW'12) (2012)
3. Rizzo, G., Troncy, R.: NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools. In: 13^{th} Conference of the European Chapter of the Association for computational Linguistics (EACL'12) (2012)
4. Steiner, T.: SemWebVid - Making Video a First Class Semantic Web Citizen and a First Class Web Bourgeois. In: 9^{th} International Semantic Web Conference (ISWC'10) (2010)
5. Waitelonis, J., Ludwig, N., Sack, H.: Use what you have: Yovisto video search engine takes a semantic turn. In: 5^{th} International Conference on Semantic and digital media technologies (SAMT'10) (2011)