Describing and Contextualizing Events in TV News Show

José Luis Redondo García
EURECOM
Biot, France
redondo@eurecom.fr

Laurens De Vocht
Ghent University - iMinds
Ghent, Belgium
laurens.devocht@ugent.be

Raphaël Troncy
EURECOM
Biot, France
raphael.troncy@eurecom.fr

Erik Mannens
Ghent University - iMinds
Ghent, Belgium
erik.mannens@ugent.be

Rik Van de Walle
Ghent University - iMinds
Ghent, Belgium
rik.vandewalle@ugent.be

ABSTRACT

Describing multimedia content in general and TV programs in particular is a hard problem. Relying on subtitles to extract named entities that can be used to index fragments of a program is a common method. However, this approach is limited to what is being said in a program and written in a subtitle, therefore lacking a broader context. Furthermore, this type of index is restricted to a flat list of entities. In this paper, we combine the power of non-structured documents with structured data coming from DBpedia to generate a much richer, context aware metadata of a TV program. We demonstrate that we can harvest a rich context by expanding an initial set of named entities detected in a TV fragment. We evaluate our approach on a TV news show.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

Media annotation; named entity recognition; entity expansion; graph search

1. INTRODUCTION

The amount of videos shared on the Web is constantly increasing. We advocate the adoption of the Linked Media principles, where video segments are annotated with structured information and linked to other video segments. A new generation of innovative video services intend to use those semantic descriptions and media fragments for providing the users a novel experience where television content and web information are seamlessly interconnected.

The annotation problem has been traditionally addressed by applying multimedia analysis techniques [1], but extracting semantic information from a video is still a challenging task. One possible approach consist in using Named Entity Recognition (NER) over the textual information attached to particular video fragment. Those techniques are an essential component within the Information Extraction field that focus on: identifying atomic information units in texts, named entities; classifying entities into predefined categories (also called context types) and linking to real world objects using web identifiers (Named Entity Disambiguation). A growing number of APIs provide such a service, like AlchemyAPI[1] or DBpedia Spotlight[2]. If the textual information attached to a video contains temporal references (e.g. subtitles), it is possible to align the entities with the time when they appear in the video. Katsiouli et al. [3] have demonstrated that applying named entity recognition techniques in combination with domain ontologies on video subtitles can produce good results for video classification.

Within the Linked Data community, a first objective is to increase the volume of interconnected data. However, from an exploitation point of view, those promising techniques still introduce some issues. On the one hand, subtitles are not always complete enough to be the only textual source to rely on. The context around a particular event is broader than what is said in a video. On the other hand, a flat list of name entities fails to characterize what is described in the multimedia content: sometimes, one also needs to know how important those entities are with respect to an event or how those entities relate to each other.

In this paper, we present an approach that generates complex structured annotations of a news event video by alleviating the lack of textual resources that limits the application of semantic extraction techniques. We first extend the initial set of descriptions about an event via Google searches and entity clustering. Secondly, we use an optimized pathfinding algorithm [2] implemented in the Everything is Connected Engine (EiCE). Applying these algorithms to Linked Data enables to resolve complex queries that involve the semantics of the relations between resources, discovering relevant resources and context-sensitive filtering resources. Each path between the resources discovered has a semantic meaning that can be traced back to the original configuration of the user and forms the basis of an explanation rather than a ranking. A major contribution in this approach is that it minimizes the size of the candidate pool of resources in order to optimize queries and increase the quality of the resulting paths.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579326.

[1] `http://www.alchemyapi.com/`
[2] `http://spotlight.dbpedia.org/`

2. APPROACH

To reconstruct the semantic context associated with one particular news video, we extract the main concepts and entities from the subtitles and explain how they are related to each other. The complete processing workflow takes as input the textual transcript of a multimedia resource illustrating an event, as well as the start and end date for which that particular event is considered.

We assume that this event has a minimal presence and coverage on the Web to ensure that the subsequent data mining techniques can collect sufficient data to reconstruct the event's context. The output of the algorithm is a pair $Context_{Event1} = [\varepsilon, \rho]$ where ε is a list of named entities together with a numeric relevance score ($\varepsilon = \{E \times \mathbb{R}\}$, E being a set of named entities classified using the NERD ontology[3]) and ρ being a set of predicates $[e_1, e_2, p]$, relating these entities ($e_1 \in E \vee e_2 \in E$).

Our hypothesis states that this representation of the events provides a sufficient source of information for satisfying the viewer's information needs and supports complex multimedia operations such as search and hyperlinking.

2.1 Named Entity Extraction

For each news item, we perform named-entity recognition over the corresponding subtitles using the NERD framework [8]. In our experiment, the language of the videos is English but NERD supports other languages. The output of this phase is a collection of entities annotated using the NERD Ontology, that comes with a first relevance score obtained from the extractors which have been used. This set includes a list of ranked entities that are explicitly mentioned during the video. Other entity based video annotation tools [5] stop at this point even when entities that can be relevant for the viewer in the context of the event are still missing. We tackle this problem by extending this first list of concepts via the entity expansion component.

2.2 Named Entity Expansion from Unstructured Resources

The set of entities obtained from a traditional named entity extraction operation is normally insufficient and incomplete for expressing the context of a news event. Sometimes, some entities spotted over a particular document are not disambiguated because the textual clues surrounding the entity are not precise enough for the name entity extractor, while in other cases, they are simply not mentioned in the transcripts while being relevant for understanding the story. This is an inherent problem in information retrieval tasks: a single description about the same resource does not necessarily summarize the whole picture.

The named entity expansion operation relies on the idea of retrieving and analyzing additional documents from the Web where the same event is also described. By increasing the size of set of documents to analyse, we increase the completeness of the context and the representativeness of the list of entities, reinforcing relevant entities and finding new ones that are potentially interesting inside the context of that news item.

The entire logic will further be described in the following subsections and mainly consist of (1) building an appropriate search query from the original set of entities, (2) retrieving additional documents about the same news event, and (3) analyzing them for providing a more complete and better ranked set of final entities, as illustrated in Figure 1.

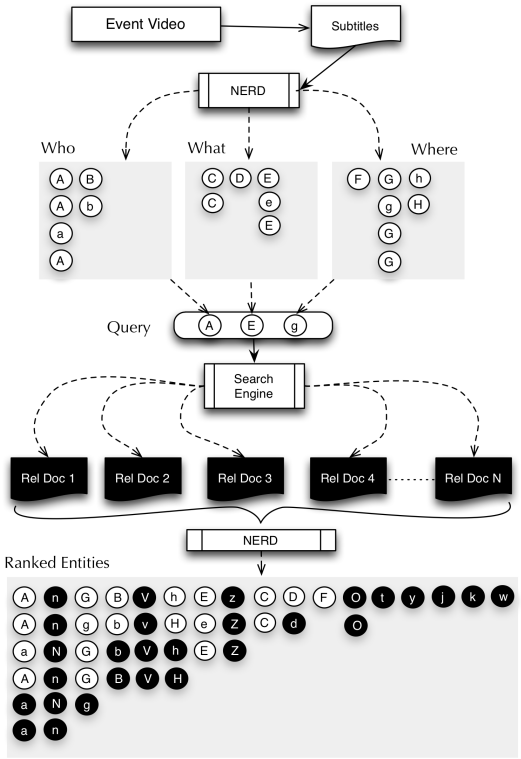

Figure 1: Schema of Named Entity Expansion Algorithm.

2.2.1 Query Generation

The *Five W's* is a popular concept of information gathering in journalistic reporting. It captures the main aspects of a story: who, when, what, where, and why [4]. We try to represent the news item in terms of four of those five W's (who is involved in the event, where the event is taking place, what the event is about, and when it has happened) in order to generate a query that retrieves documents associated to the same event.

First, the original entities are mapped to the NERD Core ontology, which considers 10 main classes: Thing, Amount, Animal, Event, Function, Organization, Location, Person, Product and Time. From those ten different categories, we generalize to three classes: the Who from `nerd:Person` and `nerd:Organization`, the Where from `nerd:Location`, and the What from the rest of NERD types after discarding `nerd:Time` and `nerd:Amount`. The When or so-called temporal dimension does not need to be computed since it is considered to be provided by the video publisher.

After generating the three sets of entities, the next step consist in ranking them in relevance according to a weighted sum of two different dimensions: their frequency in the transcripts and their former relevance scores coming from the named entity extractors. We have defined the function *filterEntities(S)* for selecting the n entities inside the set of entities S whose relative relevance

$$R_{rel}\left(e_i, S\right) = R\left(e_i\right) / Avg\left(R\left(e_i\right)\right) \quad (1)$$

falls into the upper quarter of the interval

[3] `http://nerd.eurecom.fr/ontology/nerd-v0.5.n3`

$$[max\left(R_{rel}\left(e_i, S\right)\right) - min\left(R_{rel}\left(e_i, S\right)\right)] \quad (2)$$

The final query is a pair

$$\text{Query}_{Event} = [\text{textQuery}, t] \quad (3)$$

where *textQuery* is the result of concatenating the labels of the most relevant entities in the sets Who, What, Where in that particular order, and t the time period dimension. This query generation is depicted in the upper part of Figure 1.

2.2.2 Document Retrieval

Once Query_{Event} is built out of the original set of named entities, it will be ready to be injected into a document search engine where additional descriptions about the news event can be found. In this situation, the kind of query generated in the previous step and the search engine chosen should be closely tied in order to maximize the quality of the obtained results. The different behavior of search engines make some alternatives more suitable than others for certain kinds of events. The way the resulting documents change in the search engines for a particular kind of event is a research question that will not be studied in this paper.

In this paper, we rely on the Google Search REST API service[4] by launching a query with the text *textQuery*. Due to quota restrictions imposed by Google, the maximum number of retrieved document is set to 30. However, as shown in the evaluation described in the Section 4, this is enough for significantly extending the initial set of entities directly spotted by NERD.

Concerning the temporal dimension, we only keep the documents published in the time period $t + t_e$. We increase the original event period in t_e because documents concerning a news event are not always published during the time of the action is taking place but some hours or days after. The value of t_e depends on many factors such as the nature of the event itself (whether it is a brief appearance in a media, or part of a longer story with more repercussion) or the kind of documents the search engine is indexing (from very deep and elaborated documents that need time to be published, to short post quickly generated by users). Based on the simple assumption that that the longer is an event, and the longer it is likely to generate buzzes, we approximated $t_e = t$ which means that we also consider document published during the course of an event.

The middle part of Figure 1 shows this process. The query is input in the search engine in order to retrieve other documents that report on the same event discussed in the original video. Those documents (colored in black in the Figure 1) will be further processed to increase the size of the collection and get additional insights about the news item.

2.2.3 Entity Clustering

In this phase, the additional documents which have just been retrieved are now processed and analyzed in order to extend and re-rank the original set of entities and consequently get a better insight about the event. Since most of the resources retrieved are Web pages, HTML tags and other annotations are removed, keeping only the main textual information. This plain text is then analyzed by the NERD framework in order to extract more named entities.

[4] `http://ajax.googleapis.com/ajax/services/search/web?v=1.0`

In order to calculate the frequency of a particular resource within the entire corpora, we group the different appearances of the same instance and check their cardinality. This is not a trivial task since the same entity can appear under different text labels, contain typos or have different disambiguation URL's pointing to the same resource. We performed a centroid-based clustering operation over the instances of the entities. We considered the centroid of a cluster as the entity with the most frequent disambiguation URL's that also have the most repeated labels. As distance metric for comparing pairs of entities, we applied strict string similarity over the URL's, and in case of mismatch, the Jaro-Winkler string distance [9] over the labels. The output of this phase is a list of clusters containing different instances of the same entity.

2.2.4 Entity Ranking

The final step of the expansion consists of ranking the different named entities obtained so far. To create this ordered list, we assigned a score to every entity according to the following features: relative frequency in the transcripts of the event video; relative frequency over the additional document; and average relevance according to the named entity extractors. The three dimensions are combined via a weighted sum where the frequency in the video subtitles has a bigger impact, followed by the frequency on the searched documents and the relevance from the extractors. The final output of the entity expansion operation is a list of entities together with their ranking score and the frequency in both the main video and in the collected documents retrieved from the search engine.

Entities with a higher $relScore_i$ in the final classification are considered more representative for describing the context than the original entities. Furthermore, we observe that:

- The bigger the sample size, and the better becomes the ranking. Entities appearing repeatedly in the additional documents will be promoted while those appearing rarely will be pushed back to the end of the list.
- Entities that originally have not been disambiguated can now have their corresponding URL if any of the similar instances appearing in the additional documents provide a link to a Web resource. The same occurs with incomplete or misspelled labels.
- Finally, some entities not spotted in the original transcripts but important in the context of the event are now included in the list of relevant items since they have been extracted from the collected documents.

2.3 Refining Event Context via DBpedia

Once the set of context relevant entities has been expanded, we will use the knowledge from structured sources to reinforce the important entities and finding relevant predicates between them.

2.3.1 Generating DBpedia paths

Before we filter the relations between resources to reinforce important entities, the candidate resources to be included in relations are pre-ranked. They are pre-ranked according to "popularity" and "rarity"which are essential components in the original PageRank algorithm [7] and which is

used to sort candidate related nodes in the EiCE. The implementation of the EiCE takes the relations into account by making use of the Jaccard coefficient to measure the dissimilarity and assign random walks based weight able to highly rank more rare resources, and guaranteeing that paths between resources prefer specific relations and not general ones [6].

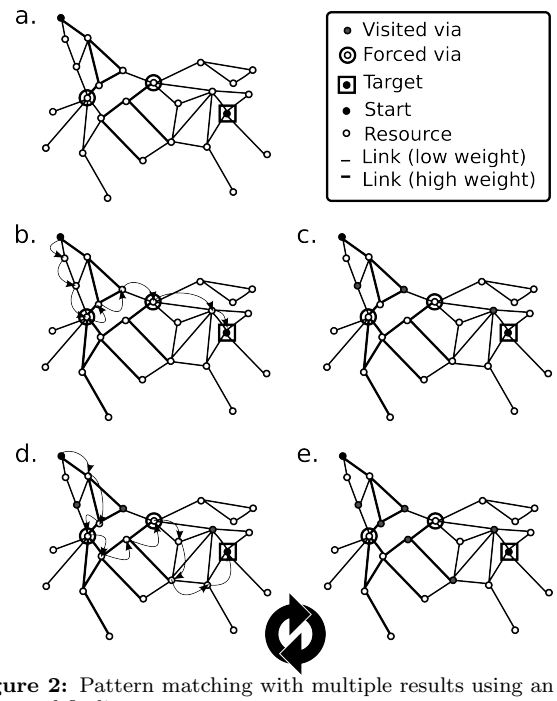

Figure 2: Pattern matching with multiple results using an iterative pathfinding process.

We pass on the main start resource, the target one and some via points. Figure 2 shows the iterative process for generating the DBpedia paths. An initial state is computed in step 2a. There are low weights and high weights. Based on the weights of the links, each path through the vias is optimized, so a path with the lowest total weight will be selected first, until the vias are added to the exclude list. The path from start to end is forced through the given via points (2b). This leads to additional visited resource as via points (2c). They occur because each computed path starts both from the start and the target resources and goes through the via points. The resources where they converge to each other are considered as new via points. These additional via points are included in the paths and therefore marked as visited. This is to make sure that in the next iteration, paths will go around the visited via (2d). The next paths are being computed over and over (2e) until a threshold number of paths is found and the context is large enough or when it takes too long to compute the next path (out of range). The final set of optimized paths is used for the context expansion.

2.3.2 Analyzing Context Relevant Paths

In this step, the DPpedia path finding technique implemented in 2.3.1 is applied over the set of entities obtained via entity expansion. Since this list is too broad, we need to establish in first place a division between what we consider the *mainEntities*(entities whose relative scores fall under the higher 25% of the relevance range), and the *additionalEntities*(the rest of entities in $E_{Expansion}$).

Afterwards, we calculate paths between all the possible pairs inside the set*mainEntities*. Once all the possible paths have been retrieved, we perform various analysis for detecting which are the most frequent classes and predicates:

- We detect the most frequent nodes ($F_{nodes} = f_{max}(n_i)$ in $Paths(mainEntities)$).
- We find the most frequent properties ($F_{prop} = f_{max}(p_i)$ in $Paths(mainEntities)$). The edges (DBpedia properties) will determine which are the most relevant properties taking part in the context of this news item.
- We study the connectivity between nodes (Adjacency Matrix $M_{i,j}$ where the distance between $e+i$ and $e+j$ is the average length of the paths linking them ($m_{i,j} = Avg(lenght(Paths(e_i, e_j))))$

The output of this phase is a re-ranked list of entities from the expansion entity set based on the paths found in DBpedia, the most important predicates and nodes inside the paths between pairs in $mainEntities$, and the adjacency matrix.

$$\{Entities_{Extension+DBpedia}, E_{Expansion}, F_{nodes}, F_{prop}, M_{i,j}\} \quad (4)$$

In the future, we plan to use the frequency measures about predicates available in F_{prop} in order to more precisely rank the named entities. By using the *commonalities* function (`http://demo.everythingisconnected.be/commonalities?between=Res1&and=Res2&allowed=property`) over the set of $mainEntities$ inside $Entities_{Extension+DBpedia}$ and for the properties with highest F_{prop}, we obtain the set of entities directly related with the original ones through the top predicates. Once more, the most frequent entities in Commonalities could promote the already existing entities in $Entities_{Extension+DBpedia}$.

3. USE CASE: SNOWDEN ASSYLUM

We have applied this method over a real scenario corresponding to the news video `http://www.bbc.co.uk/news/world-europe-23339199`, from the BBC News Europe Web site. The main story behind this media content is the request of asylum made by Edward Snowden to the Russian government. In an airport in Moscow, he publicly express his desire to obtain political help while he can find a safe way to reach the Latin American countries that offer a safe harbor. The time period for which this particular event is relevant goes from 2013-07-06 to 2013-07-17.

3.1 Named Entity Extraction

In a first step, Name Entity Extraction techniques are applied over the video transcript using NERD. We show below the list of entities directly extracted via this procedure, which brings a first approximation to the context of the news media item.

3.2 Named Entity Expansion

In a second step, the original set of entities is expanded as described in Section 2.2. The query generated for retrieving additional documents has the text "Edward Snowden asylum Russia", and is bounded to the period from 2013-07-06 to 2013-07-28. After analyzing the additional documents

Label	Relevance	Sentiment	Type	URI
Russia	0.809216	Mixed	Location	DBpedia:Russia
Edward Snowden	0.717369	Mixed	Person	DBpedia:/Edward_Snowden
South America	0.56586	Mixed	Location	DBpedia:South_America
president Putin	0.459811	positive	Person	DBpedia:Vladimir_Putin
president	0.401138	negative	JobTitle	DBpedia:President
Moscow	0.352101	Mixed	City	DBpedia:Moscow
CIA	0.334887	neutral	Organization	DBpedia:CIA
Bolivia	0.324607	neutral	Location	DBpedia:Bolivia
Obama	0.321901	negative	Person	DBpedia:Barack_Obama

Table 1: Raw result from Named Entity Extraction with NERD

Label	Relevance	F_{video}	F_{Docs}	Type	URI
Russia	1.0	7	264	Location	DBpedia:Russia
Edward Snowden	0.80479	2	227	Person	http://dbpedia.org/resource/Edward_Snowden
US	0.61643	5	160	Location	DBpedia:United_States
Vladimir Putin	0.39383	1	111	Person	DBpedia:Vladimir_Putin
asylum	0.32876	4	80	Thing	DBpedia:Right_of_asylum
Barack Obama	0.31506	1	88	Person	DBpedia:Barack_Obama
Moscow, Russia	0.30479	1	85	Location	DBpedia:Moscow
American president	0.19178	2	48	Thing	DBpedia:President_of_the_United_States
Central Intelligence Agency	0.19178	0	56	Location	DBpedia:Central_Intelligence_Agency
Anatoly Kucherena	0.147260	0	43	Person	-
extradition	0.116438	2	26	Thing	DBpedia:Extradition
White House	0.10616	0	31	Location	DBpedia:White_House
Sheremetyevo	0.0890	0	26	Location	DBpedia:Sheremetyevo_International_Airport
WikiLeaks	0.08219	0	24	Organization	DBpedia:WikiLeaks
Washington's	0.075342	0	22	Location	DBpedia:Washington,_D.C.

Table 2: Top entities obtained via Named Entity Expansion

Nodes		Properties	
URL	Frequence	URL	Frequence
DBpedia:Wilmington,_North_Carolina	14	http://dbpedia.org/ontology/country	44
DBpedia:United_States	38	http://dbpedia.org/ontology/birthPlace	64
DBpedia:Russia	12	http://purl.org/dc/terms/spatial	42
DBpedia:conference/AIPR/2008	11	http://dbpedia.org/ontology/almaMater	26
DBpedia:Washington,_D.C.	10	http://dbpedia.org/ontology/location	21
DBpedia:Igor_Panarin	8	http://dbpedia.org/ontology/profession	20
DBpedia:Chap_Petersen	8	http://dbpedia.org/ontology/nationality	14
DBpedia:North_Carolina	8	http://dbpedia.org/property/leaderTitle	14
DBpedia:Independent_(politician)	8	http://dbpedia.org/ontology/occupation	14

Table 3: Top middle nodes and properties inside DBpedia paths

Label	Rel_{Final}	$Rel_{extension}$	$Conectivity$	$F_{inPaths}$	URI
Russia	1.0	1.0	6.10	12	DBpedia:Russia
Edward Snowden	0.79469	0.80479	7.60	-	http://dbpedia.org/resource/Edward_Snowden
US	0.77188	0.61643	9.40	38	DBpedia:United_States
Vladimir Putin	0.38383	0.39383	5.50	-	DBpedia:Vladimir_Putin
Barack Obama	0.37231	0.31506	6.69	-	DBpedia:Barack_Obama
Washington's	0.21544	0.07532	6.10	10	DBpedia:Washington,_D.C.
Moscow, Russia	0.35991	0.30479	6.90	-	DBpedia:Moscow
asylum	0.25036	0.32876	1.70	-	DBpedia:Right_of_asylum
vv American president	0.21571	0.19178	6.60	-	DBpedia:President_of_the_United_States
White House	0.12096	0.10616	6.30	-	DBpedia:White_House
Sheremetyevo	0.11905	0.0890	3.90	-	DBpedia:Sheremetyevo_International_Airport
Central Intelligence Agency	0.09171	0.19178	2.80	-	DBpedia:Central_Intelligence_Agency
Anatoly Kucherena	0.08780	0.147260	0.00	-	-
extradition	0.06693	0.116438	0.00	-	DBpedia:Extradition
WikiLeaks	0.04432	0.08219	0.00	-	DBpedia:WikiLeaks

Table 4: Top entities re-ranked via DBpedia clues

retrieved by the search engine, the instances are re-ranked according to their global frequency.

The final result after the expansion operation is a bigger list of entities (more than 100). In some cases, the previously spotted entities (like *Extradition*) have been promoted in the hierarchy while others (like *Right_of_asylum*) have been discovered in the new documents. For this use case, we have taken only a fixed number of top entities ($mainEntities = 15$). Those entities will be used as input to the third step of context refinement.

3.3 Refining Event Context via DBpedia

Given the set of *mainEntities* obtained from the Named Entity Expansion operation, we study how close two graph nodes e_i, and e_i are connected according to the normalized average length of all the intermediate paths between them in DBPedia. The results are expressed in the form of an adjacency matrix, which allows to visually analyze the connections between pairs of entities inside *mainEntities*.

$$M_{i,j} = \begin{pmatrix} - & 0.4 & 0.6 & 1.0 & 0 & 0.2 & 1.0 & 0.6 & 0 & 0 & 0 & 0.7 & 1.0 & 0 & 0 \\ 0.4 & - & 0.9 & 0.6 & 1.0 & 0.8 & 0.1 & 0.8 & 0.4 & 0 & 0 & 0.7 & 0.6 & 0 & 0.7 \\ 0.6 & 0.9 & - & 0.8 & 0.7 & 1.0 & 0.5 & 1.0 & 0.8 & 0 & 0 & 0.9 & 0.5 & 0 & 0.9 \\ 1.0 & 0.6 & 0.8 & - & 0 & 0.4 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0 & 0.2 \\ 0 & 1.0 & 0.7 & 0 & - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2 & 0.8 & 1.0 & 0.4 & 0 & - & 0.3 & 0.9 & 0.3 & 0 & 0 & 1.0 & 0 & 0 & 0.8 \\ 1.0 & 0.1 & 0.5 & 0.9 & 0 & 0.3 & - & 0.8 & 0.4 & 0 & 0 & 0.7 & 0.9 & 0 & 0.7 \\ 0.6 & 0.8 & 1.0 & 0 & 0 & 0.9 & 0.8 & - & 0 & 0 & 0 & 1.0 & 0 & 0 & 0.8 \\ 0 & 0.4 & 0.8 & 0 & 0 & 0.3 & 0.4 & 0 & - & 0 & 0 & 0 & 0 & 0 & 0.3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & - & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & - & 0 & 0 & 0 & 0 \\ 0.7 & 0.7 & 0.9 & 0 & 0 & 1.0 & 0.7 & 1.0 & 0 & 0 & 0 & - & 0 & 0 & 0.7 \\ 1.0 & 0.6 & 0.5 & 0.9 & 0 & 0 & 0.9 & 0 & 0 & 0 & 0 & 0 & - & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & - & 0 \\ 0.4 & 0.7 & 0.9 & 0.2 & 0 & 0.8 & 0.7 & 0.8 & 0.3 & 0 & 0 & 0.7 & 0 & 0 & - \end{pmatrix} \quad (5)$$

This matrix allows to identify different aspects and insights about the *mainEntities* analyzed. First, the upper left quarter of the matrix, corresponding to the connectivity between the top 8 entities from $E_{Expansion}$ is more dense in connections, which means that the concepts are semantically closer. This corroborates the results obtained in the previous steps since it reinforces the idea of a coherence in the proposed context. At the same time, other entities such as *WikiLeaks* have a lower number of connections to the rest of elements in the context and are pushed back inside the *mainEntities* set. For the same reason, an entity such as *Sheremetyevo* is promoted given its relationship with entities like *Russia*, *Moscow*, or *Vladimir Putin*. However, this logic can lead to wrong inferences if applied on entities where the disambiguation URL's is missing, like *Anatoly Kucherena*, but this problem lies in the absence of some DPpedia resources and falls apart the scope of our approach. This intuitive measure, notated in Table 4 as *Conectivity*, is calculated as the sum of the distances from one entity to the rest. This score is then a weighted factor over the former expansion relevance in order to re-rank the entities.

In addition, we study the frequency of the different middle nodes and properties inside the set of items that conform the paths for calculating $M_{i,j}$. The most repeated ones are listed in Table 3. In the case of the nodes, we can observe various interesting findings: on the one hand, we get the location *Wilmington*, the city where Edward Snowden was born and where he still has family, and important Russian-American political figures like *Chap Petersen* or *Igor Panarin*. On the other hand, some entities already spotted in the expansion phase are reinforced, like for example *Russia* or *Washington D.C.* Finally, the absolute frequency of properties reveals also meaningful insights: the first three predicates, and also the fifth and the seventh are clearly related to the spatial dimension, while the forth, sixth and eighth are related with political matters, which clearly makes sense since an asylum request implies a territorial and political conflict.

The *Conectivity* scores are combined together with the frequency of intermediate nodes in the paths ($F_{inPaths}$) and the former relevance indexes from the entity extension phase ($Rel_{extension}$) for obtaining a new ranked list of entities that intends to better represent the context of the news event.

4. EVALUATION

We have evaluated the result obtained by our approach against a set of entities provided by an expert. We plan to extend this evaluation to a larger corpora and include a more exhaustive evaluation in future work. We start from the list of entities provided by the expert, together with some insights about the relevance of the concept inside the scope of the video:

We first evaluate the set of entities spotted by the expert against traditional Named Entity Extraction results from NERD. About precision and recall, 3 entities have been spotted out of the 7 while in total 12 were proposed as candidate results. The lawyer of the case, Anatoly Kucherena, is never mentioned in the subtitles, so it has not been detected. Even if the word airport is present in the subtitles, we are not sure which one is relevant given that there are two airport in Moscow.

Next, we evaluate against the set of *MainEntities* inside $E_{Expansion}$. In this case, the precision and recall indexes are better: there are 6 entities spotted, out of 7. In total, the 15 entities inside *MainEntities* were offered as output. Asylum, probably one of the more representative concepts behind this news item, has been correctly detected and disambiguated. *Anatoly Kucherena*, the lawyer involved in the defense of Edward *Snowden*, has been correctly proposed in the result, even if no disambiguation URL is provided. The name of the airport is now known: *Sheremetyevo* and it has been correctly disambiguated. Other important entities such as *human-rights* and *extradition* are now present in the final set, helping to complete the context of the news.

Finally, we compare the expert's feedback against the final DBpedia $E_{Expansion+DBpedia}$ ranked set of entities. There are no differences for the spotted *MainEntities*. However, we observe that other improvements have been achieved concerning the ranking order inside *MainEntities*. For example, the entity "Sheremetyevo" is now scored higher, which makes sense from the event's perspective since it is the most specific location where the action takes place. Some other relevant entities include the prime Minister of Russia, Igor Shuvalov, which was not detected via entity expansion. They have not been included in the set of results but we will study this possibility for future developments. Also, the adjacency matrix opens room for many other context analysis like topic clustering, which could help to give further insights about an event.

5. CONCLUSIONS

We presented an approach for context-aware annotating news events, designed to precisely harvest program descriptions starting from named entities recognized in TV video transcripts. Because the entities initially spotted are typically insufficient for covering the broader range of concepts that best describe a particular news clip, we expanded this set by analyzing additional textual documents about the same event.

The flat ranked list of concepts is afterward completed with cues about their connectivity obtained by analyzing the DBpedia paths that exist between them. In particular, we promoted entities which are better interlinked inside the context and with higher frequency in the nodes of the generated DBpedia paths. Finally, we identify the important predicates linking those entities. This leads to a more accurate re-ranking of the entities belonging to this news event.

The preliminary results indicate that we can successfully expand the initial set of recognized entities with more relevant concepts not detected by pure named entity recognition approaches. Exploring DBpedia paths along the named entities occurring in news media leads to a more accurate ranking of important concepts and even if it does not necessary introduce any new top item, it brings forward more related entities with additional information about the broader context of an event.

Our future work includes a better analysis of the DBpedia properties for detecting the most likely used predicates between context entities. By prioritizing paths using those properties and following them, we should be able to find other relevant entities more easily or to justify the ones that have been previously selected. The adjacency matrix we generate will be also used for more in depth analysis like topic extraction based on entity clustering. Finally, we will extend the evaluation to other use cases to see if the contributions exposed in this paper can be confirmed with other scenarios.

Entity	Comments
Edward Joseph Snowden	Public figure. He is the "who" of the news. The subject of the main sentence
Rusia	The location, but also an actor, the indirect object of the main sentence "to whom".
Political Asylum	This is related to the "what" of the news. This is the Snowden's request, the direct object.
CIA	Background information on related to Snowden, since he is an ex-CIA employee. An axe in a wider sense, not this item in particular, but SnowdenÖs history.
Sheremetyevo Airport	Specific location of the news.
Anatoly Kucherena	Secondary actor and speaker in the video. Information about an interview or person expressing his opinion.
US Department of State	Involved organization. Not mentioned but related to speaker.

Table 5: Relevant Entities from an Expert

Acknowledgments

The authors would like to thank Michiel Hildebrand and Lilia Pérez Romero from CWI Amsterdam for providing us with ideas and use cases for illustrating the approach described in this paper. This research has been partially funded by Ghent University, and the European Union's 7th Framework Programme via the project LinkedTV (GA 287911).

6. REFERENCES

[1] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra. Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*, 51(1):279–302, 2011.

[2] L. De Vocht, S. Coppens, R. Verborgh, M. Vander Sande, E. Mannens, and R. Van de Walle. Discovering meaningful connections between resources in the Web of Data. In *6^{th} Internation Workshop on Linked Data on the Web (LDOW'13)*, 2013.

[3] P. Katsiouli, V. Tsetsos, and S. Hadjiefthymiades. Semantic video classification based on subtitles and domain terminologies. In *1^{st} International Workshop on Knowledge Acquisition from Multimedia Content (KAMC'07)*, 2007.

[4] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *IEEE 11^{th} International Conference on Computer Vision (ICCV'07)*, pages 1–8, 2007.

[5] Y. Li, G. Rizzo, J. L. Redondo Garcia, and R. Troncy. Enriching media fragments with named entities for video classification. In *1^{st} Worldwide Web Workshop on Linked Media (LiME'13)*, Rio de Janeiro, Brazil, 2013.

[6] J. L. Moore, F. Steinke, and V. Tresp. A novel metric for information retrieval in semantic networks. In *The Semantic Web: ESWC 2011 Workshops*, pages 65–79, 2012.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the web. Technical report, Stanford InfoLab, 1999.

[8] G. Rizzo and R. Troncy. 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)*, Avignon, France, 2012.

[9] W. E. Winkler. Overview of record linkage and current research directions. In *Bureau of the Census*, 2006.