The Concentric Nature of News Semantic Snapshots

Knowledge Extraction for Semantic Annotation of News Items

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

Giuseppe Rizzo
EURECOM
Biot, France
giuseppe.rizzo@eurecom.fr

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

ABSTRACT

The Web enables to have access to silo-ed information describing news articles, often offering a multitude of viewpoints that, once combined, can provide a broader picture of the story being reported on the news. In this paper, we propose an approach that automatically extracts representative features of a news item, namely named entities, from textual content attached to a video item (subtitles) and from a set of documents from the Web collected using entity expansion techniques. Approaches relying on entity expansion generally try to collect and process the important facts behinds a particular news item, but they are often too dependent on frequency-based functions and information retrieval techniques thus neglecting the multi-dimensional relationships that are established among the entities. We propose a concentric-based approach that enables to represent the context of a news item, by harmonizing into a single model the representative entities, which can be extracted using information retrieval and natural language processing techniques (*Core*), and other entities that get prominent according to different dimensions such as informativeness, semantic connectivity, or popularity (*Crust*). We compare our approach with a baseline by analyzing the compactness of the generated summary on an existing gold standard available on the Web. Results of the experiments show that our approach converges faster to the ideal compact news snapshot with an improvement of 30.1% over the baseline.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*retrieval models, search process*

Keywords

Semantic Annotation, Entity Expansion, News Item

1. INTRODUCTION

A single presentation of a news item taken individually generally fails to illustrate the complexity of the event being reported. The Web has offered a data space where one can find different content such as citizen-based blogs, journalistic articles or social media posts teaming up for generating a multi-rich ecosystem of complementary news content.

The intuitive hypothesis that we formulate is that a single news item is often not enough to capture the complete story being reported, because it generally assumes the viewer has prior knowledge of the event, and is thus incomplete, biased or even partially wrong if interpreted in isolation. The most common strategy to complement the description of a given news item is to enrich the original content with additional data collected from external sources. Research initiatives have approached this problem by exploiting the absolute frequency of entities that can be extracted from additional but related documents. However, this results in large amounts of unreliable and repeated information, leaving to the user the burden of processing this potentially related data to build an understanding of the event. Redondo *et al.* [10] propose to use entity expansion techniques over the open Web to produce a ranked list of entities in order to generate a so-called Newscast Semantic Snapshot (NSS), which complements the initial set of detected entities in subtitles with other item-related entities captured from Web documents. The ranking is achieved using a set of functions that measure the absolute entity frequency across the collected document, expert rules, and global popularity of the entity. Largely inspired by the aforementioned work, we performed an experimental and critical assessment of this method, observing that the frequency functions and pure information retrieval techniques that were used during the NSS generation neglect the intrinsic relationships that the entities hold. In particular, we observed that frequency-based rankings and their hybrid approaches are appropriate for spotting essential or inherent entities, but they fail to consider other entities that are still important for understanding the context of a news item due to other reasons such as interestingness, informativeness or popularity.

In this paper, we recast the problem of generating a NSS by exploiting and harmonizing in a single model different semantic relationships established between the news' entities. Instead of tackling the problem from a pure list-based oriented model, where all the different news related phenomena are projected into a single dimension, we propose a concentric-based approach with two main layers called **Core** and **Crust**. This knowledge representation model better supports the complex and multi-dimensional relations es-

tablished among the entities involved in a news item and allows to formalize the distinction between the representative entities, which better characterize the essence of the news item, and the relevant ones, that are potentially interesting because of different reasons that link them to the Core. This graph-based knowledge representation considers the multidimensional nature of those relationships, allowing us to focus in different desired features for the final NSS, like representativeness and compactness. The final ranking order is then delegated to the final applications that display the data and inevitably project the rich spectrum of relationships among entities describing an event into a single and easier to consume dimension.

The remainder of this paper is organized as follows: Section 2 presents some related work. Section 3 presents a critical assessment of our previous NSS approach and enables to list the key ideas of the work presented in this paper. We then state the hypothesis in Section 4. In Section 5, we describe our approach that leads to the concentric model creation. We propose an evaluation in the Section 6. Finally, we summarize our main findings and outline some future work in Section 7.

2. RELATED WORK

The need of a NSS for feeding certain applications is already a concept investigated in some research works and prototypes like [9], which have probed the benefit for users when browsing the "surrounding context" of the newscasts. Research efforts have underlined the importance in professional journalism of proper annotations to illustrate news event. Projects like NEWs have studied how to disambiguate named entity in the news domain by continuously learning over news streams [4]. In the domain of Social Networks, named entities are used for identifying and modeling events and detecting breaking news. Entities have been also exploited in video classification when the textual information attached to a video contains temporal references (e.g. subtitles) [5]. The need for contextualizing news related documents has already been addressed in some research works such as [1], where the authors describe the challenge of generating textual summaries of news items based on an entity-centric process of traversing a graph. This generates a subgraph containing the most relevant entities and the structural relationships among them. In the particular case of this paper, those relationships are used as the means for building human readable summaries instead of aiming to create general and shareable data structures like the NSS is.

In some cases, like [13] and [12], this contextualization process is not based on entities but on other items like other documents or text snippets. However, a machine readable representation of the context of the news items is still needed. In particular, the context is built up from different keywords directly mentioned in the news item with no use of external sources in the middle. Another difference is the existence of human intervention during the workflow since the keywords have to be manually highlighted by the users.

Other examples of news contextualization efforts are found in [14], where the authors aim to complement the original news document with the most interesting reactions from different online platforms. In order to identify relevant candidates, they propose to look at different dimensions including interestingness, informativeness, opinionatedness, and popularity. The motivations for implementing a concentric model approach presented in this paper actually includes the need of dealing with all those different dimensions.

To the best of our knowledge, the only related work in the news domain that has been carried out grounding the power of enriching the set of initial entities by using an entity expansion algorithm are [8] and [10]. The former includes a naive document collection strategy for bringing in related documents, and proposes a pure frequency-based algorithm for ranking entities that is complemented by looking into the DBpedia knowledge base for ensuring the coherency of the final list of entities via the number and length of paths among them. The later improves the former in several directions: document retrieval mechanism, semantic annotation, and creation of the NSS. However, it fails in exploiting the semantic relationships between the spotted entities, projecting different relevancy dimensions into a single ranking list, and ignoring other important aspect of the generated context like representativeness and compactness.

3. MOTIVATION

This section summarizes the different research efforts made for critically assessing and extending the experiments described in [10].

3.1 Improving the NSS Generation Baseline

In this section, we try to extend the strategy described in [10] by further exploiting additional relevance indicators that could have been missing during our previous study, with the objective of improving the Average Normalized Discounted Cumulative Gain ($MNDCG$) at N. This measure [2] considers different levels of relevance and gives more priority to items ranked at top positions since they are more likely to be examined by a user. Some changes that brought some improvement over the original approach are:

- Exploit Google relevance: Documents obtained from Google Custom Search Engine (CSE)[1] come ordered, so the ones on the top are potentially more relevant, and therefore related, to the studied news item. Assuming that entities spotted within those higher ranked documents are more important than the ones found in less interesting documents, we can weight them differently when summing up scores in frequency functions (see Section 4.1 in [10]). This adaptation enables to gain 1.8% in $MNDCG$ over the best configuration from the original approach.

- Promote subtitle entities: Entities detected in subtitles can be better considered since they are explicitly mentioned in the video speech and therefore, are more likely to be relevant to what is being reported. By analyzing different ratios for weighting subtitle entities versus related document's ones, the combination (1:4) brought the best outcome, obtaining a percentage increase of 2.5% of $MNDCG$.

Some attempts that did not improve or even slightly reduced $MNDCG$ are:

- Exploit Named Entity Extractor's confidence: Similarly to the Google relevance, confidence scores produced by the entity annotators [11] can be used to differently ranked entities when accumulating them on frequency. This new

[1] `https://www.google.com/cse/all`

dimension brought a percentage decrease in $MNDCG$ of 0.2%.

- Interpret popularity dimension: Candidate entities proposed by the popularity function (see Section 4.3 in [10]) need to be combined together with the outcome of the frequency measures to provide a single ranked list of entities. Scores coming from both dimensions were simply summed in [10]. In order to go a step further, we have created a function $F : R_{Pop} \rightarrow R_{Freq}$ which linearly transforms scores produced by the popularity function into values inside the range of the frequency functions R_{Freq} before performing the addition. Unfortunately, this lead to a percentage decrease of 1.4% in $MNDCG$.

- Perform the clustering before entity filtering: By filtering entities first, we get rid of many noisy annotations and partially correct results that can contaminate the generation of the NSS. It is possible to proceed the other way around: run the clustering operation over the whole set of annotations and then filter out clusters according to their entity centroids. Intuitively, the clustering phase can benefit from partially correct annotations since they could still balance clusters by becoming part of some of them. However, the results following this modified workflow are less performing than the former approach because the filtering stage becomes too aggressive by removing some important entity clusters that have low representative centroids (we observed a percentage decrease of 0.60% in $MNDCG$).

According to the experiments conducted, the situation did not bring a significant improvement in the original scores. A deeper study of the results reveals that prioritizing certain ranking dimensions quickly brings valid results but also discards relevant entities that were selected before. The complete workflow is too dependent on frequency functions and pure information retrieval techniques thus neglecting the semantic relationships present in the ideal NSS of a news event. In fact, frequency driven rankings and their hybrid approaches are appropriate for spotting essential or inherent entities, which must be inside the final result, but are failing to consider other entities that model more specific details about the news item that are relevant in the NSS creation for particular reasons, but barely mentioned in the related documents.

3.2 Thinking Outside the Box

Given the limitations found in the approach described in [10], we try to tackle the problem from a different angle and reconsider the conditions that a NSS should match in order to be properly consumed by other users and applications.

After studying that state-of-the-art ranking algorithms have reached a ceiling in performance in terms of $MNDCG$, the first question to be answered is if the data retrieved via the collection phase can still offer some rooms for improvement. By looking at further positions ($10 < n < 50$) in the ranking generated by the best run at 10 in [10] ($Best_{MNDCG_{10}}$ = {L1+Google, 2Weeks, NoSchema, F3, Freq, ExpertRules}[2]) we can plot the score evolution when considering bigger NSS. The curve in Figure 1 suggests that cumulative gain keeps increasing at bigger n values even if the gradient is not pronounced.

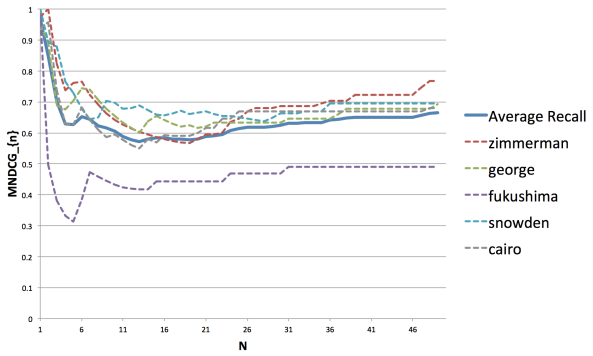

Figure 1: $MNDCG_{1-50}$ for run $Best_{MNDCG_{10}}$ in [10].

Cumulative gain is a measure that weights more the matches in the top ranks. In order to make this analysis more ranking agnostic in Figure 2, we analyzed the recall R when the size of the NSS goes till the position 50. The slope is now steeper and clearly reflects that more relevant entities are still found at lower positions in the ranking and can potentially be moved to the top.

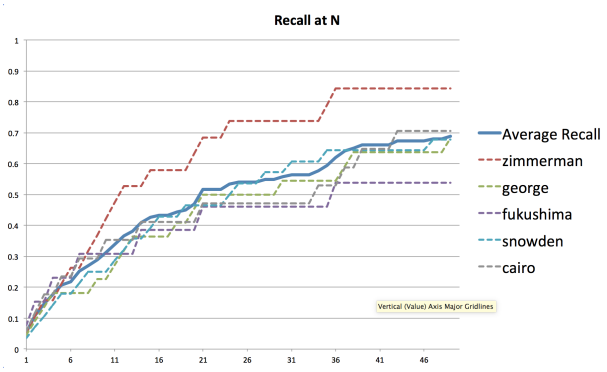

Figure 2: $Recall_{1-50}$ for $Best_{MNDCG_{10}}$ = in [10].

To check how good $Best_{MNDCG_{10}}$ performs compared to other configurations in terms of Recall, the complete set of configurations has been re-run in order to see which one is working the best. The top 10 configurations, labeled from $Ex1$ to $Ex10$ are shown in Table 1.

Table 1: Executed runs and their configuration settings, ranked by R_{50}.

Run	Collection			Filter	Ranking			Result
	Sources	T_W	Schema.org		Freq	Pop	Exp	R_{50}
Ex1	L2+Google	1W		F3	Gaussian	✓	✓	0.7224
Ex2	L2+Google	1W		F3	FreqGoogle		✓	0.7119
Ex3	L2+Google	1W		F3	Freq	✓	✓	0.7115
Ex4	L2+Google	1W		F3	FreqGoogle	✓	✓	0.707
Ex5	L2+Google	1W		F3	Gaussian	✓		0.707
Ex6	L1+Google	1W		F3	Gaussian	✓	✓	0.7031
Ex7	L2+Google	1W		F3	Gaussian			0.6944
Ex8	L2+Google	1W		F3	Freq			0.693
Ex9	L1+Google	2W		F3	Gaussian	✓	✓	0.6919
Ex10	Google	2W		F3	Gaussian		✓	0.6908

We first observe that $Best_{MNDCG_{10}} \neq Best_{R_{50}}$. In other words, at further positions in the ranking, other strategies

[2]L1 is a white list of three news web sites (The Guardian, New York Times, Al Jazeera), L2 is a white list of 10 news web sites, filtering and ranking functions are defined in [10].

bring up a higher amount of relevant entities, even if they were not that well performing in ranking at $n <= 10$. Additionally, the popularity dimension, which was getting down cumulative gain scores, seems to bring up relevant entities at higher n values according to the configuration of the four best performing runs. In conclusion, there is room for improvement, but also a need for:

- Changing the knowledge acquisition method behind the approach. As already explained before, pure information retrieval techniques are not enough to explain the relevancy of an entity for a particular news item. Certain entities are important for summarizing the context of the news items, while others are informative for users who just want to discover something beyond the obvious facts. Dimensions like popularity or semantic relatedness can not be projected into a single ranking dimension. Instead, they need to be equally considered when promoting important entities so they can take part on the NSS.
- Changing the final objective. The task of bringing as many relevant entities as possible inside the NSS will be prioritized against being too precise in ranking. This is in line with the idea of generating a flexible and application-independent NSS, which intends to be comprehensive enough to contain as many entities as possible for better representing the context of a news item. Our outcome aims to produce a solid semantic representation that can properly feed different prototypes and tools with information needs that vary.

4. HYPOTHESIS

This section formally describes the problem we are addressing and the main hypothesis we are formulating. In a nutshell, the semantic snapshot of a news item (NSS) can be modeled following a schema of concentric entity layers. This kind of representation helps to better reproduce the context of a news event and ease the task of identifying the different relevant entities for various dimensions.

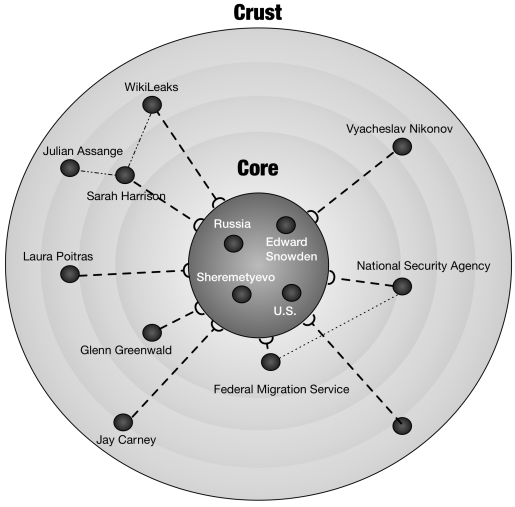

Figure 3: Concentricity of the news item *"Fugitive Edward Snowden applies for asylum in Russia"*

In this model, we are considering two main entity layers that can annotate a news item: **Core** and **Crust**.

Core: It is composed of a small number of key entities which are essential to identify an event. Those entities have the highest degree of representativeness and can better summarize the main facts attached to the event. They are frequently mentioned in related documents and are therefore spottable via frequency-based functions. Sometimes, they are too obvious for the user, but they are the key elements for describing the facts. They are *semantically compact* in the sense that it exists numerous semantic relationships between each entity. Let $E_{d_i} = e_{1_{d_i}}, ..., e_{n_{d_i}}$ be the bag of entities belonging to the related document d_i, and E_T the union of all E_{d_i}:

$$Core = G\left(e_1, e_2, ...e_c, e_i\right), e_i \in E_{d_i} \quad (1)$$

$$\text{(Frequency Prominence)} \forall e_i \in Core, f(e_i) > t \mid 0 \ll t < 1 \quad (1a)$$

$$\text{(Compact)} \forall e_i, e_j \in Core, S(e_i, e_j) > s_{core} \mid 0 \ll s_{core} \leqslant 1 \quad (1b)$$

Crust: It is composed of a larger number of entities that describe particular details of a news items. Those entities are mentioned in some specific related documents, but they are not always spottable via frequency-based measures. They are not necessarily pairwise related (not semantically compact). Their relevancy is instead grounded on the existence of special relations (including popularity, serendipity, etc.) between those entities and the $Core$.

$$Crust = G\left(e_1, e_2, ...e_c\right), e_i \in E_{d_i} \quad (2)$$

$$\text{(Core Attached)} \forall e_i \in Crust, S_(e_i, Core) > s \mid 0 \leqslant s \ll 1 \quad (2a)$$

Those two layers can be aggregated into a single structure in order to build the so called **News Semantic Context (NSS)**. In our hypothesis, and differently than in other research works such as [10], this structure will be a graph of entities since the relationships established among the elements inside the NSS are more important than their absolute ranking, therefore providing a data structure that remains as flexible as possible.

$$SSN_{concentric} = Core \oplus Crust \quad (3)$$

Given this new nature characterizing a NSS, maximizing cumulative gain is not a priority for this study. Therefore, we need to define a new and more ranking-agnostic **Objective Function**. One possibility is measuring the recall R. However, this metric does not consider different degrees of entity relevance which are available in the ground truth (see Section 6). In order to exploit this additional information, an extended recall index R^* has been defined. This measure takes into account the different scores of the relevant entities and gives a more accurate idea about the coverage provided by a particular NSS:

$$R^*(\text{NSS}) = \frac{\sum Score_{GT}(e_{NSS_i})}{\sum Score_{GT}(e_{g_{t_i}})} \quad (4)$$

Furthermore, this research aims to tackle the problem from a bigger perspective by studying a wider region of the entity annotations spectrum and not only focusing in a particular NSS at n. Let us define Res_{Ap} as a list of N entities produced by a certain approach Ap, we define a Semantic Snapshot NSS_{Ap} as the n first entities in Res_{Ap}.

$$NSS_{Ap} = \{e_0, e_1, ..., e_n\} \mid n < \left|Res_{Ap}\right| = N \quad (5)$$

We introduce the so-called compactness Com of an entity set Res, given a certain function f and a value v:

$$Com\left(R, f, v\right) = |min(NSS \in Res)| \mid f(NSS) \geq v \quad (6)$$

This measure helps to indicate if a particular set of entities is able to produce concise NSS while still keeping the goal of $f(NSS) > v$. Our objective is to produce entity sets with minimum compactness Com. The function R^* is used to quantify the coverage.

Having defined the different concepts and dimensions above, and being the results Res_{Exp} and Res_{Conc} sets of N relevant entities about a particular news item generated via our former implementation and the concentric model approach respectively, we formulate the following **Hypothesis**:

$$Com\left(Res_{Exp}, R^*, R^*(Res_{Exp})\right) > Com\left(Res_{Conc}, R^*, R^*(Res_{Exp})\right) \tag{7}$$

According to this hypothesis, the concentric model approach has to be able to produce more concise, cleaner, and potentially easier to consume New Semantic Snapshots than the related research efforts implementing an unidimensional ranking.

5. THE APPROACH

This section presents our proposed approach for generating News Semantic Snapshots based in a concentric model. The workflow is composed of the following steps (Figure 4): after executing state-of-the art entity expansion and ranking strategies with the best configurations possible to bring to the top positions as many relevant semantic annotations as possible (see the grey part on the left side, labeled as (**1**), we build the concentric model of the news items in tree different steps: generating the $Core$, the $Crust$ and the final NSS, as depicted in the white part on the right side (**2**).

Named Entity Expansion and Ranking. The first step consists of executing the expansion approaches presented in [10] for generating a list of entities Res_{Exp}. The objective is to reduce the size of the spectrum of annotations to be considered while building the concentric model and, therefore, relaxing the complexity of having to work over the entire set of entities. Taking as input the metadata that news broadcasters offer about the items they publish, the query $q = [\mathrm{h}, \mathrm{t}]$ is built, where h is the video headline and t is the publication date. This query allows us to collect a set of event-related documents D from the open Web over which the semantic annotation process is performed. After removing HTML tags and other markup annotations, the feature space is then reduced and each document d_i is represented by a bag of entities $E_{d_i} = e_{1_{d_i}}, ..., e_{n_{d_i}}$, where each entity is defined as a triplet $(surface\_form, type, link)$. A filtering process prepares entities to be clustered applying a centroid-based algorithm based on strict string similarity over the $link$ and $surface\_form$. The output of this phase E'_{d_i} is further processed to promote the named entities that are highly related to the underlined event, based on entity appearance in documents, popularity peak analysis and domain experts' rules in order to produce a ranked list of entities Res_{Exp}, which feed the concentric NSS generation approach.

Core Generation. The Core generation process works over the set of filtered entity annotations per document E'_{d_i} in order to identify the entities with the higher level of representativeness for a particular event. As stated in our hypothesis, we exploit the frequency prominence principle expressed in Definition 1a to spot the candidates. In particular, the absolute frequency of an entity within the set of retrieved documents D, noted as $f_a(e_i, D)$, and the Bernoulli appearance rate across all documents $f_{doc}(e_i, D)$ will be considered according to the following formula:

$$f_{Core}(e, D) = f_{doc}(e_i, D) + \frac{f_a(e_i, D)}{f_{doc}(e_i, D)} \tag{8}$$

After ordering the entities according to $f_{Core}(e, D)$, top ranked entity start to be added in the $Core$ until we found one which is not semantically connected to *all* the other ones already included. This way, we ensure the second condition for the $Core$ generation expressed in Equation 1b, the semantic compactness. In order to check if an entity e_i is connected with other e_j, we identify existing paths between them in a particular knowledge base KB. The process of detecting those paths enables to identify other resources $r \in KB$ that can materialize such connections. Those intermediate resources are promoted via dimensions such as "popularity" and "rarity" that are essential components in the original PageRank algorithm [7]. The implementation makes use of the Jaccard coefficient to measure the dissimilarity and assign random walks based weights that are able to highly rank those rare resources, guaranteeing that paths between resources promote specific relations other general ones [6]. Assuming there is a number p of paths between the entities e_i and e_j ($path_{i,j}$) and being $|path_{i,j}|$ a path length as number of links among resources r, we define the similarity function S_{KB} as:

$$S_{KB}(e_i, e_j) = \sum_{1}^{p} \frac{1}{|path_{i,j}|} \tag{9}$$

As stated in Equation 1b, two entities are considered well-connected if $S_{KB}(e_i, e_j) > s$.

Crust generation. Taking again as input the results Res_{Exp}, we use different similarity functions working in certain relevancy dimensions in order to detect which entities are next to the $Core$ (as shown in Definition 2, $S_{(e_i,Core)}$). Therefore, the $Core$ acts like the contextual anchor where $Crust$ candidate entities are attached to. In the current approach, two functions grounded on different principles have been considered:

- The semantic relationships between resources in knowledge bases, via the number and length of paths between an entity $e_i \in Crust$ and the entities in the $Core$. Based on the definition of $S_{KB}(e_i, e_j)$ in Equation 9, we define the similarity function $S^*_{KB}(e_i, Core)$ as the sum of the different similarities between e_i and $ej \in Core$:

$$S^*_{KB}(e_i, Core) = \sum S_{KB}(e_i, e_j) \mid e_j \in Core \tag{10}$$

- The number of web documents talking simultaneously about a particular entity e_i and the $Core$. This function, noted as $S_{Web}(e_i, Core)$, identifies documents in the Web talking about a candidate entity and the $Core$ at the same time, while keeping in mind the original volume of documents containing them separately. Let E be a set of entities and the function $hits_s(E)$, the number of web documents where all $e_i \in E$ are mentioned, $S_{Web}(e_i, Core)$ is:

  1. Directly proportional to the square of the number of pages talking about the $Core$ and the candidate entity at the same time $hits_s(e_i + Core)$.
  2. Inversely proportional to the number of pages talking about the $Core$ ($hits_s(Core)$).

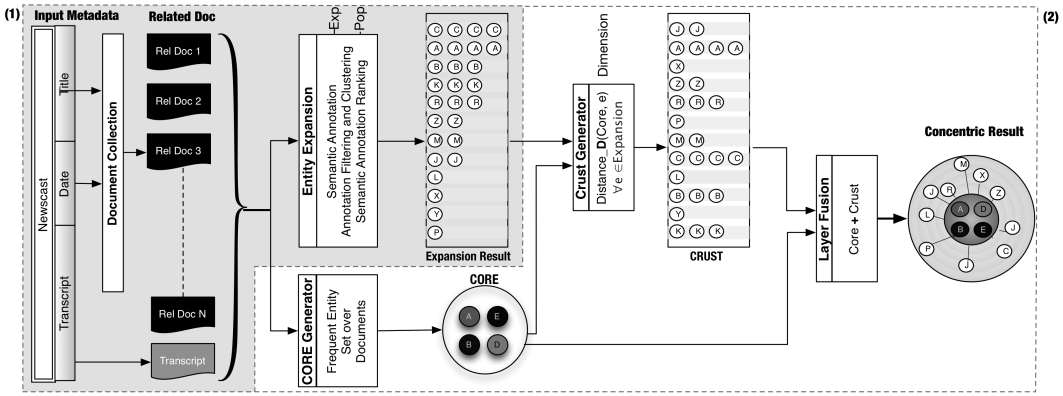

Figure 4: Concentric-based approach for generating News Semantic Snapshot using Named Entity Expansion

3. Inversely proportional to the number of pages talking about the candidate entity alone ($hits_s(e)$). If the entity was already highly mentioned all over the Web, like in the example of very famous persons, the volume of documents mentioning that entity together with the $Core$ has to be also big enough in order to be considered.

$$S_{Web}(e_i, Core) = \frac{hits_s(Core + e_i)^2}{hits_s(Core) * hits_s(e_i)} \quad (11)$$

The different similarity functions helps to populate the $Crust$ with the first top c entities spotted via each dimension.

NSS Generation. In a last step, the entities coming from the $Crust$ are attached to the $Core$ via the scores produced by the similarity functions described above in order to generate the final NSS of a news item. At this stage, the result is a graph of different event related entities. To evaluate this approach, different projection functions $f : G \rightarrow L$ have been created.

6. EVALUATION

This section describes the experimental settings and the results of the concentric model approach against a gold standard. We re-used a dataset composed of 5 ranked list of named entities that each semantically annotate one video item. Named entities are extracted from the subtitles, video image, text contained in the video, articles related to the subject of the video and event suggested by a journalist expert. After building a candidate set of entities, this set was presented to 50 participants via an online survey that were asked to rate their level of interestingness. The methodology for building this dataset is thoroughly described at `https://github.com/jluisred/NewsEntities`, where links to the list of entities and scores per video are also available.

6.1 Experimental Settings

This section explains the configuration settings used during the execution of the experiments in order to produce the concentric model based news annotations Res_{Conc}. The first important parameter to be considered is the length of the entity spectrum that will be analyzed. In order to go beyond the $n = 10$ used by [10] and to target the positions studied in the Section 3.2, the experiments have been configured to work over the first 50 entities coming from the previous expansion phase.

Named Entity Expansion and Ranking. In order to identify which entity expansion configuration can potentially serve as the best basis for generating NSS, we have studied the values of R^*_{50} over the complete set of runs considered in [10]. Table 2 shows the top 8 configurations:

Run	Collection			Filter	Ranking			Result
	Sources	T_W	Schema.org		Freq	Pop	Exp	R^*_{50}
Ex1	L2+Google	1W		F3	Gaussian	✓	✓	0.755
Ex2	L2+Google	1W		F3	Freq	✓	✓	0.7532
Ex3	L2+Google	1W		F3	Gaussian	✓		0.7457
Ex4	Google	2W		F3	Gaussian		✓	0.745
Ex5	L2+Google	1W		F3	FreqGoogle		✓	0.7448
Ex6	L2+Google	1W		F3	FreqGoogle	✓	✓	0.7424
Ex7	L1+Google	1W		F3	Gaussian	✓	✓	0.7346
Ex8	L2+Google	1W		F3	Gaussian			0.7333

Table 2: Expansion runs ranked by R^*_{50}

We have selected the first two runs in Table 2 as candidates for feeding the concentric model approach:

- First run $Ex1$ uses the second whitelist and Google, F3 filtering, one week temporal window, Gaussian function, expert rules and popularity. This configuration is also the top result when ranking by R_{50}.
- Second run $Ex2$ uses the second whitelist and Google, F3 filtering, one week temporal window, absolute frequency ranking function, expert rules and popularity. This configuration is also third in R_{50} and fourth in $MNDCG_{50}$. It is to be noted that restricting web sites to the ones embedding Schema.org markup lead to poor performance when one wants to maximize the recall of named entities.

CORE Generation. Entities have been ranked according to the frequency based Function 8. In order to perform the clustering operation that explores the frequency of the entities, we have considered two different strategies: $Core1$, based on Jaro-Winkler string distance [15] over the $surface\_form$, and $Core2$ based on exact string matching of the $link$.

As entities in the core usually express the most general, upper-level concepts that drive the story behind the news item, we will use DBpedia in order to discover relationships between the candidate entities via the similarity Function 9. In particular, we have used the optimized path-finding algorithm [3] implemented in the Everything is Connected Engine (EiCE). During the process of filtering only the n-top

semantically well-connected entities, two entities have been considered as properly connected if there is at least a path of length 5 between them $S_{KB}(e_i, e_j) > \frac{1}{5} = 0.20 = t$

CRUST generation. For generating the Crust, the two Functions 10 and 11 have been considered with the following parameters:

- $S^*_{KB}(e_i, Core)$ has been configured to work over DBpedia in order to find connectivity between the entities and the $Core$. However, this general purpose knowledge base does not work well with the more fine-grained entities available in the $Crust$, even after relaxing the threshold t in the function $S_{KB}(e_i, e_j)$ for also considering paths of length up to 10. We have empirically detected many missing relations among some entities that, according to the story being told in the news item, should be connected each other. For future work, we plan to rely in other news domain specific dataset potentially containing more links about the studied event.

- $S_{Web}(e_i, Core)$ has been configured to work over an instance of a Google Custom Search Engine where $hits_s(e)$ is the number of documents retrieved. In particular, we have set up the engine with no particular sites to crawl, no temporal filtering and the English language.

After discarding the first similarity function, there will be only one possible configuration for this phase.

NSS Generation. In order to project the final graph-based structure into a list that could be evaluated against the Ground Truth, two possible approaches have been taken into account:

- $Core + Crust$: entities in the $Core$ are placed at the top positions of the result list Res_{Conc}. Entities in the $Crust$ are added just after those.

- $CrustBased$: all entities in the $Crust$ are added to the list of results Res_{Conc}. We also calculate $S_{Web}(e_i, Core)$ for the entities in the $Core$ and we place them in the right position according to this similarity score.

6.2 Result

Initially, we have performed a specific evaluation of the $Core$, consisting in calculating the precision P of the entities contained in this layer. Results are close to 95%, which means that the great majority of the $Core$ entities are in the ground truth, underling their importance as the driving force to build the NSS of a news item.

Afterward, we have executed the concentric model approach with the different configurations selected in the experimental settings (a total of $2 * 2 * 1 * 2 = 8$). We have also consider two baselines produced by traditional entity expansion techniques. They are identified by the run names **BAS01** and **BAS02** respectively. In addition, we have contemplated the existence of an ideal system able to generate the same perfect ranking available in the ground truth, in order to understand how good we could potentially get. Assuming $R^*(Ex1) = 0.755 \approx 0.753 = R^*(Ex2)$, in Table 3, we order the results in terms of compactness
$Com\,(Cmn, R^*, R^*(Exn))$ for each of the 6 concentric model configurations, breaking down the scores by video (v_1, v_2, v_3, v_4, v_5), and showing the final average in the right column in bold.

We can first observe that the average compactness of the concentric model approach is smaller than the original ones represented by the baselines $BAS01$ and $BAS02$ for all 5 videos. This proves our original hypothesis:
$Com\,(BAS01|BAS02) > Com\,(\text{Cm0-7})$. Additionally, if we compare the best concentric model run $Cm4$ with the best baseline $BAS01$, and having as ideal objective the compactness of $IdealGT$, we can report a percentage increase of 30.1% over a traditional entity expansion method.

In order to see in a more intuitive way how better is the evolution of R^* values over the whole spectrum, we plot the scores from the concentric based run $Cm0$ against its baseline $BAS01$ (Figure 5).

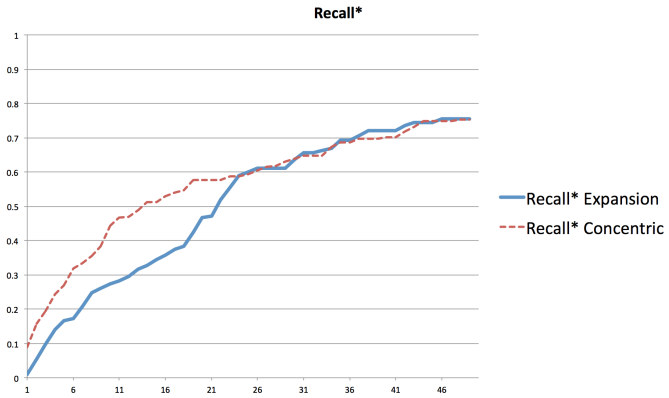

Figure 5: $R^*_{1-50}(Res_{Exp})$ vs. $Recall^*_{1-50}(Res_{Conc})$: the concentric model approach gets faster to higher values of R^*

Approximately from $n = 0$ to $n = 22$, we can see how dashed line Res_{Conc} gets faster to higher values of R^*, which mean, it can potentially produce more representative NSSs from lower n positions in the obtained results.

7. CONCLUSION

The different applications and tools consuming information about news can benefit from the computation of a News Semantic Snapshot (NSS), which summarizes and explicitly describes the context of a news items. For generating such a data structure, we can not rely exclusively on the metadata of that particular news item. Instead, there is a need of relying on the Web and to complement the available information via a process called named entity expansion. However, this step brings in numerous non-relevant entities that need to be discarded. One way of promoting news related entities is to rely on pure information retrieval methods, but those approaches are not suitable for spotting other relevant entities that are linked to the main story according to aspects such as popularity, serendipity, or semantic proximity. For overcoming those difficulties and be able to exploit the entity semantic relations, we have proposed a concentric based model for generating the NSS. This model proposes two layers: the $Core$, composed of the most representative entities, which are well-connected between them and spottable via frequency measures, and the $Crust$, which sometimes includes unfrequent entities that are attached to the $Core$ via particular similarity functions. In order to ensure the semantic compactness of the $Core$, we have looked at the existence of DBpedia paths between each entity pair. For establishing connections between the $Core$ and the entities in the $Crust$, we observe that a general purpose knowledge base such as DBpedia is not necessarily ideal due to the fine grained na-

Run	Expansion				$Com\,(R, f, v)$					
	Collection	Core	Crust	Fusion	v_1	v_2	v_3	v_4	v_5	Avg
IdealGT		-	-	-	16	11	22	27	19	**19**
Cm4	Ex2	CoreA	S_{Google}	Core_Crust	21	9	41	44	45	**32**
Cm5	Ex2	CoreA	S_{Google}	CrustBased	20	14	41	44	45	**32.8**
Cm6	Ex2	CoreB	S_{Google}	Core_Crust	27	10	43	44	42	**33.2**
Cm0	Ex1	CoreA	S_{Google}	Core_Crust	22	13	42	43	47	**33.4**
Cm1	Ex1	CoreA	S_{Google}	CrustBased	21	16	42	43	47	**33.8**
Cm7	Ex2	CoreB	S_{Google}	CrustBased	27	13	43	44	42	**33.8**
Cm2	Ex1	CoreB	S_{Google}	Core_Crust	28	13	43	43	44	**34.2**
Cm3	Ex1	CoreB	S_{Google}	CrustBased	28	16	43	43	44	**34.8**
BAS01	L2+AllGoogle, 1W F3 Gaussian + EXP + POP	-	-	-	41	45	34	41	37	**39.6**
BAS02	L2+AllGoogle, 1W F3 Freq + EXP + POP	-	-	-	24	39	49	48	39	**39.8**

Table 3: Compactness of concentric model results VS compactness of baselines and ideal ground-truth-based result set

ture of the entities in the *Crust*. However, other dimensions like the web presence of both the *Core* and *Crust* entities have successfully highlighted relationships promoting relevant entities. The experiments in terms of R^* over a set of results produced by the concentric model approach have revealed a significant improvement in the level of compactness of the new method compare with traditional expansion methods, which allows to produce more concise and at the same time representative NSS.

Our future work includes: *i)* extending the amount of news items (videos) in the current ground truth; *ii)* increasing the length of the spectrum of annotations used for feeding up the concentric model up to the complete list of annotations (which is harder due to quota restrictions in some services such as Google CSE); *iii)* being able to spot not only the degree of connectivity between the entities in the *Crust* and the *Core* but also the predicates that characterize those connections; and *iv)* studying the role of the model in tracking the evolution of news events over the time, where we expect that the *Core* remains more or less stable, and the entities in the *Crust* will vary according to the particular moment of the event.

Acknowledgments

This work has been partially supported by Bpifrance within the NexGen-TV Project, under grant number F1504054U, and by the European Union's 7th Framework Programme via the project LinkedTV (GA 287911).

8. REFERENCES

[1] S. Chhabra. Entity-centric Summarization: Generating Text Summaries for Graph Snippets. In *23rd ACM International Conference on World Wide Web (WWW)*, pages 33–38, Seoul, Korea, 2014.

[2] W. B. Croft, D. Metzler, and T. Strohman. *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2010.

[3] 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 *6th International Workshop on Linked Data on the Web (LDOW)*, 2013.

[4] N. Fernandez, J. A. Fisteus, L. Sanchez, and G. Lopez. IdentityRank: Named entity disambiguation in the news domain. *Expert Systems with Applications*, 39(10):9207–9221, 2012.

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

[6] J. L. Moore, F. Steinke, and V. Tresp. A novel metric for information retrieval in semantic networks. In *3rd International Workshop on Inductive Reasoning and Machine Learning for the Semantic Web (IRMLeS)*, pages 65–79, 2011.

[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] J. L. Redondo García, L. De Vocht, R. Troncy, E. Mannens, and R. Van de Walle. Describing and Contextualizing Events in TV News Show. In *2nd International Workshop on Social News on the Web (SNOW)*, pages 759–764, Seoul, Korea, 2014.

[9] J. L. Redondo García, M. Hildebrand, L. Romero, and R. Troncy. Augmenting TV Newscasts via Entity Expansion. In *11th Extended Semantic Web Conference (ESWC)*, pages 472–476, 2014.

[10] J. L. Redondo García, G. Rizzo, L. Perez Romero, M. Hildebrand, and R. Troncy. Generating the Semantic Snapshot of Newscasts using Entity Expansion. In *15th International Conference on Web Engineering (ICWE)*, 2015.

[11] G. Rizzo, M. van Erp, and R. Troncy. Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web. In *9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 2014.

[12] N. K. Tran, A. Ceroni, N. Kanhabua, and C. Niederée. Back to the Past: Supporting Interpretations of Forgotten Stories by Time-aware Re-Contextualization. In *8th ACM International Conference on Web Search and Data Mining*, pages 339–348, Shanghai, China, 2015.

[13] N. K. Tran, A. Ceroni, N. Kanhabua, and C. Niederée. Time-travel Translator: Automatically Contextualizing News Articles. In *24th ACM International World Wide Web Conference (WWW)*, pages 247–250, Florence, Italy, 2015.

[14] T. Štajner, B. Thomee, A.-M. Popescu, M. Pennacchiotti, and A. Jaimes. Automatic Selection of Social Media Responses to News. In *19th ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 50–58, 2013.

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