Capturing News Stories Once, Retelling a Thousand Ways

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

We live in a constantly evolving world where news stories and relevant facts are happening every moment. For each of those stories, numerous news articles, posts, and social media reactions are created, offering a multitude of viewpoints about what is happening around us. Many applications have tried to deal with this complexity from very different angles, targeting particular needs, reconstructing certain parts of the story, and exploiting certain visualization paradigms. In this paper, we identify those challenges and study how an adequate news story representation can effectively support the different phases of the news consumption process. We propose an innovative model called News Semantic Snapshot (NSS) that is designed to capture the entire context of a news item. This model can feed very different applications assisting the users before, during, and after the news story consumption. It formalizes a duality in the news annotations that distinguishes between *representative* entities and *relevant* entities, and considers different relevancy dimensions that are incorporated into the model in the form of concentric layers. Finally, we analyze the impact of this NSS on existing prototypes and how it can support future ones.

Categories and Subject Descriptors

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

Keywords

News Stories, News Semantic Snapshot, Entity Annotation

1. INTRODUCTION

Even the a-priori conventional stories that we daily consume have some underlying facts that, during certain situations and for some particular users, become important and need to be unveiled. Those facts can be described in very different ways: publishers may aim to emphasize certain aspects of the story, target a specific audience, or respond to particular viewer's needs. The role of the consumers can also evolve over the time, from a passive and less engaging behavior to a deep-into-the-details mode that requires deeper knowledge about the facts being reported. In this highly challenging ecosystem where stories are spread all over different pieces of information, interpreted by many different users and presented by various data sources, the existence of a model representing the entire context of the news item becomes highly relevant.

The construction of such advanced story representation has already been addressed in [5]. In our previous work, we made the hypothesis that a single video news item is often not enough to capture the complete story being reported and can be biased or even partially wrong. We have developed various information retrieval techniques for combining the original news content with additional data collected from other external sources. This process, called Named Entity Expansion, is able to produce a ranked list of named entities that complements the initial set of detected entities in video subtitles with other item-related entities captured from Web documents. The top n items in this list build the conceptual structure called the Newscast Semantic Snapshot (NSS) of a news story. In [2], this NSS evolves from a plain ordered list of entities to a multi-layered concentric model, which is more appropriate for representing the duality between the most representative entities and the other ones that are relevant to the context of the news item due to diverse reasons such as interestingness, informativeness or popularity.

In this paper, we analyze how this NSS can support the different requirements derived from the news consumption process. This structure needs to (1) be easy to exploit and flexible enough for giving an answer to different applications, (2) deal with the duality present in the news annotations, by differentiating between entities that better summarize a story and the ones that acquire relevancy as the story is further consumed, and (3) emphasize the relationships established between the different entities inside the context of the story, focusing more in the reasons for having such a connection and less in their absolute importance inside the story. In the last part of the paper, we will analyze some prototypes that project the rich spectrum of relationships within entities into a simpler and human easier way to consume the story, in order to understand how they can benefit from such a news item representation.

K-CAP 2015, October 07-10, 2015, Palisades, NY, USA
 ISBN 978-1-4503-3849-3/15/10 ...$15.00
DOI: http://dx.doi.org/10.1145/2815833.2816951.

2. A MODEL FOR NEWS CONSUMPTION: THE NEWS SEMANTIC SNAPSHOT

The News Semantic Snapshot (NSS) is a graph structure that tries to represent the entire context of a news story, where nodes are named entities and edges represent relationships among them. According to the hypothesis stated in [2], a NSS can be modeled following a schema of concentric entity layers. The NSS aims to exploit and harmonize

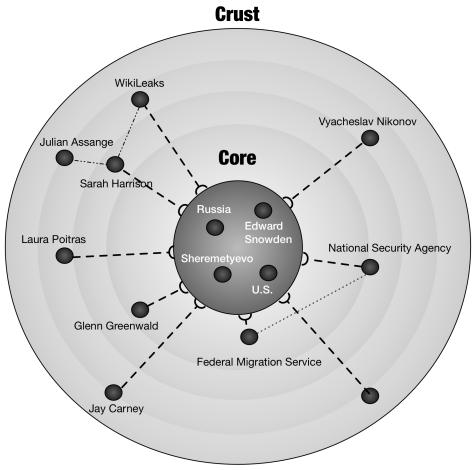

Figure 1: Concentricity of the news item "Fugitive Edward Snowden applies for asylum in Russia"

in a single conceptual model different semantic relationships established between the news' entities. This model, based on the findings made in [2], makes explicit a duality in the entities via two main layers namely the **Core** and the **Crust**. The former is composed of a small number of key entities that are essential to identify a story. Those entities have the highest potential to better summarize the main facts behind the news story. They are frequently mentioned in related documents and therefore spottable via frequency-based functions. In Figure 1, the *Core* is composed of the entities Russia, Snowden, U.S., and Sheremetyevo (the airport where the action is taking place), which are the seeds for a good understanding of the story. On the other hand, the *Crust* is composed of the entities expressing the particular details around the news items. They are mentioned in some specific related documents, but they are not always spottable via frequency-based measures. Their relevancy is instead grounded on the existence of relations such as popularity, serendipity among those entities and the *Core*. In Figure 1, some entities such as Anatoli Kucherena (Edward Snowden's lawyer), are not so prominent at a first glance, but they definitely play a role in the story and can contribute to a better understanding of the facts. The semantic context of a news item can be therefore built by combining the *Core* and the *Crust* into a single data structure $NSS_{concentric} = Core \oplus Crust$. In order to go deeper in the formalization of the NSS, we focus on some other aspects of this conceptual model that are important in the news consumption scenario.

Reconciling Relevancy Dimensions. The concept of relevancy is extremely wide and complex. It depends on several variables that two different persons that ought to judge the relevancy of an entity would rarely agree on. However, the layer based representation used in the NSS better supports the complex and multi-dimensional relevancy relations established among the entities involved in a news item and allows to formalize the potential reasons that are linking them to the *Core*. The *Crust* becomes then a place for hosting different relevancy dimensions [8], which bring diversity to story description: entities denoting opinions, informativeness, serendipity, popularity, interestingness, unexpectedness.

Finding Predicates to Entity Relations. In [2], we discussed the importance of discovering and explicitly establishing relations among the entities inside the NSS. We propose now a step further by considering, not only the unlabeled relations, but also explicit predicates characterizing the entity links. Finding such property names and formalizing those entity dependencies is still an open challenge. In [2], co-occurrences of entities in documents collected from the Web revealed how tight were the relations in the context of the story. A further analysis of those documents could help to provide additional information enabling to label the predicates.

As the NSS is a graph-based structure, this information can be straightforwardly incorporated into the model. The predicates established between the elements inside can become labeled links thanks to the flexible nature of the model. Prototypes can exploit such kind of annotations in order to make the users aware of the reasons that make those entities relevant.

Tracking Stories over Time. In most of the cases, the different facts, which shape the story plot, happen in a chronological order. This implies that the entities and predicates involved in such facts are especially relevant during a particular period of time within the time span of the story. Once more, the flexible graph-based NSS model can support edges annotated with temporal references in order to reflect when a particular entity plays an important role in the plot of the story or holds a specific relationship with others. Tracking the evolution of those relationships in time opens a room for timeline based summarization prototypes that highlight the milestones characterizing the story evolution.

3. THE NEWS CONSUMPTION PARADIGM

Potentially, there are numerous ways of consuming a news story St_{News}. Each tool or application displaying information about a news item follows a different philosophy, targets a different audience, and presents the main facts from a different angle. Despite this variety of alternatives, in this section, we propose a model for classifying those news consumption approaches. Having for reference the time, the user is actually consuming the news document d_{news} describing the story St_{News}, we have identified three main phases: the $before$, the $during$, and the $after$. In each of those phases, a user's behavior is different: there is an evolution in the understanding of the story and consequently in the information requirements of the applications presenting the story.

3.1 The before

Users, in this phase, have normally not consumed the main news item yet, so their understanding about the story is limited. Most of the times, they require a quick and easy way to interpret the main facts so they get to know in a glance what

the news is talking about. In other cases, additional content is displayed in order to illustrate the news context. This type of recommendations look for content that is very similar to d_{news}, leaving diversity aside. A special application category under this phase includes the advanced summaries that aim to fully tell the story without having consumed the original content.

3.2 The during

It corresponds to the time the user is watching the main document illustrating the story, d_{news}. It normally implies a passive information activity where the users are pretty much focused in the task of consuming the document without engaging in any other actions. During this phase, the user's knowledge grows from the background information provided in the $before$ phase to a most detailed understanding of the news. A good example of prototypes under this category is a second screen application aiming to illustrate what is being said on the news with minimal user interaction.

3.3 The after

The user became fully aware of the basics of the story and wants to go deeper into the details, switching to an active mode: browsing description of entities categorized into dimensions or ultimately jumping to other related stories. The level of interaction drastically increases since the user becomes more engaged, moved by the curiosity of discovering more details. The main document d_{news} can be enriched with additional content, focusing on diversity, and detailing some specific facts of St_{News}. Other applications falling under the same consumption phase are advanced interactive summaries with browsing capabilities.

3.4 The NNS in the Consumption Process

We formulate, as hypothesis, that the NSS of a news item is a knowledge representation model that effectively captures the context of a story and can support different existing news applications. This graph layer-based structure helps to populate very diverse prototypes aiming to support users in interpreting the news. For the sake of illustrating our hypothesis and without claiming to provide an exhaustive plot based on quantitative data, Figure 2 shows how the duality between $Core$ and $Crust$ in the concentric model can better satisfy the evolution of user consumption needs across the different consumption phases, as a result of changes in aspects like viewers' knowledge about the story, user's engagement and content diversity.

As defined in the Formula 1, Section 4 in [2], entities in the $Core$ usually express the most general, upper-level concepts that drive the story behind the news item. Even if those entities will be present in almost every stage of the news consumption, they have a stronger decisive role in the $before$ phase (left side of Figure 2), decreasing in importance as d_{news} is consumed and the $after$ visualizations come into play.

The $during$ is a special phase that requires both $Core$ and $Crust$ entities (middle part of Figure 2), specially when they are mentioned in the document d_{news}. In particular, $Core$ entities start to be less demanded since they have often been consumed during the $before$, while $Crust$ entities start bringing added value in revealing the non-obvious facts around the story plot.

In the last phase of the consumption process, the so-called

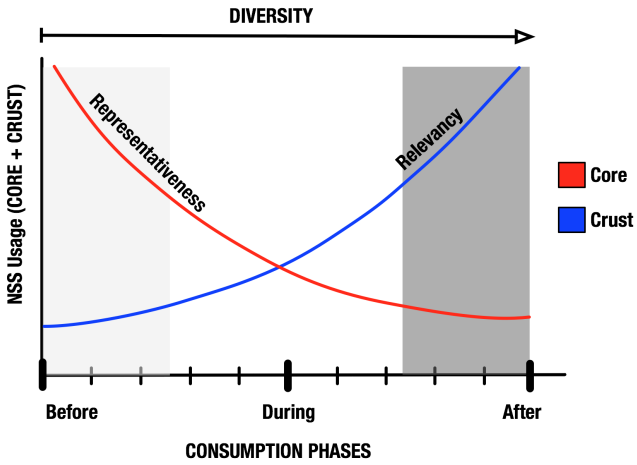

Figure 2: $Core$ and $Crust$ usage along the different consumption phases

$after$, users have already a fair understanding of the news story. The entities in the $Core$, which were highly present during the previous phases, become often too obvious and are therefore not so critical to be used. Instead, the entities in the $Crust$ bring those particular details that users want to consume, in an attempt to move from a general understanding of the news item to explore specific story details (right side of Figure 2). Since the $Crust$ considers different relevancy dimensions, applications can easily move along them and bring the diversity desired at those latest stages of the consumption process. In addition, for those applications displaying timeline based summaries, we propose an additional hypothesis stating that the $Core$ remains stable in time and have less interest, while the $Crust$ contains the entities that bring the stand-out information in particular periods of time and need to be displayed.

4. AN ECOSYSTEM OF NEWS APPLICATIONS

In this section, we review existing applications and prototypes for consuming news. We classify them according to the different consumption phases identified in Section 3, and we analyze how they would benefit from a graph representation model like the News Semantic Snapshot in order to make a first qualitative evaluation of our hypothesis.

Prototypes for Before Consuming the News Item. In [6], we presented an approach for getting a quick overview of a video content enabling the user to decide if he is interested or not in the story. We automatically select some fragments within the video called Hotspots, which contain annotations with higher frequency scores. This notion of representativeness is clearly aligned with the definition of the $Core$. Entities inside this layer could be straightforwardly used for the Hotspots creation.

Something similar occurs in [1] where some small video fragments are hyperlinked to others based on some visual or topical similarity. Even if such a task can be tackled using multimodal analysis techniques, applications at early stages of the news consumption focus mostly in finding content as similar as possible to the original. Therefore, entities in the $Core$ are suitable for triggering searches on document

indexes where that additional content can be found.

Prototypes During the Consumption of the News Item. In [4], we presented a second screen application implementing a slideshow that gives the user access to factual information about Person-type, Location-type and Organization-type entities that are related to a news story displayed in the main screen. *Core*-like entities are still shown when mentioned in the video to illustrate the story as a whole, but there is an increasing use of other entities clarifying more specific facts in the video as they are displayed (*Crust*-like entities). Given the absence of a NSS in [4] to feed such prototype, we developed a first implementation of the Entity Expansion Algorithm [5] in order to recreate a very basic version of the *Crust*.

Prototypes for After Consuming the News Item. We find an example of an application illustrating the story after consuming the news item in [5]. The active mode of the demo targeted the idea of a user who wants to further dig into the details of the story via some additional content that is proposed along different dimensions. Some of those additional content facets can easily match the layers envisioned in the *Crust* definition, like *Opinions from Experts* that aligns to *opinions*, or *In Other Sources* that aligns to *informativeness*, revealing the importance of the multi-layer philosophy inside the News Semantic Snapshot.

We can also include applications that offer advanced interactive visualizations for summarizing the entire context of the story, an extremely challenging task, even when performed by experts in the domain. The representation model offered by the NSS considers relations between entities that can help to implement conceptual diagrams where entities are related to each other via particular connections that can be visualized like in [7].

A third example in this category are the time-based representations that break down the story in relevant facts that are chronologically represented over the X axis. In [3], we analyzed the story of the Italian Elections 2013 during one week after the voting process. One of the main difficulties encountered during its implementation was to filter out those entities that were buzzing during the entire week so they became obvious like *Italy*, and to promote instead those entities that peaked in relevance during certain moments of the week for particular reasons, like *Merkel* who had a meeting with the Italian president on the 1st of March. This phenomena reinforces our hypothesis about the *Core* entities remaining stable in time and becoming useless for such timelines prototypes, while the ones in the *Crust* bringing the interesting facts that need to be shown.

Other Prototypes. In addition to the previous prototypes, there are also other applications targeting a bigger portion of the news consumption phases spectrum. A good example can be found in the online demo Hyperted[1] that offers an innovative way to consume TED talks, by supporting the user not only at the pre-consumption stage via Hotspots calculation, but also the during the viewing through entity highlighting and in the period just after by linking to similar courses and other related TED chapters.

5. CONCLUSION

The different applications consuming news can benefit from the existence of a graph-based model able to capture the entire context of the story. We proposed a model called the News Semantic Snapshot that is grounded on the existence of semantic relations between entities describing a news story. This model has a concentric nature considering two main layers: the *Core*, which includes the most representative and frequent entities, and the *Crust* which is composed of additional entities that become relevant because of certain relationships happening between them and the *Core*. In addition, we have identified three phases of the news consumption process: the before, the during, and the after. The NSS can support applications falling inside each of those categories. As support of the underlined hypothesis, we have reviewed some of our past prototypes, identifying the challenges we faced when capturing the story information, and explaining how the use of NSS to feed their needs would have led to a better implementation.

In future work, we plan to experiment with new techniques for automatically populating the model. We aim to use existing domain specific knowledge bases for further contextualizing the relationships present in the NSS, and exhaustively evaluate what is the relevancy of the various dimensions that composed the *Crust* so that users get the best consuming experience possible.

[1] `http://linkedtv.eurecom.fr/Hyperted/`

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).

6. REFERENCES

[1] E. Apostolidis, V. Mezaris, M. Sahuguet, and et. al. Automatic fine-grained hyperlinking of videos within a closed collection using scene segmentation. In *22^{nd} ACM International Conference on Multimedia*, Orlando, USA, 2014.

[2] J. L. R. García, G. Rizzo, and R. Troncy. The Concentric Nature of News Semantic Snapshots. In *8^{th} international Conference on Knowledge Capture (KCAP)*, 2015.

[3] V. Milicic, J. García, G. Rizzo, and R. Troncy. Tracking and Analyzing the 2013 Italian Election. In *10^{th} Extended Semantic Web Conference (ESWC), Demo Track*, pages 258–262, 2013.

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

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

[6] J. L. Redondo Garcia, M. Sabatino, P. Lisena, and R. Troncy. Finding and sharing hot spots in Web videos. In *13^{th} International Semantic Web Conference (ISWC), Demo Track*, 2014.

[7] G. Rizzo, T. Steiner, R. Troncy, R. Verborgh, J. L. Redondo García, and R. Van de Walle. What Fresh Media Are You Looking for? Retrieving Media Items from Multiple Social Networks. In *1^{st} International Workshop on Socially-aware Multimedia (SAM)*, pages 15–20, Nara, Japan, 2012.

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