Generating Semantic Snapshots of Newscasts using Entity Expansion

José Luis Redondo García[1], Giuseppe Rizzo[1], Lilia Perez Romero[2], Michiel Hildebrand[2], Raphaël Troncy[1]

[1] EURECOM, Sophia Antipolis, France,
{redondo, giuseppe.rizzo, raphael.troncy}@eurecom.fr
[2] CWI, Amsterdam, The Netherlands,
{L.Perez, M.Hildebrand}@cwi.nl

Abstract. Television newscasts generally report about the latest event-related facts occurring in the world. *Per se* a newscast delivers partial information thus neglecting the whole picture of the event that is often assumed as known. Relying exclusively on the broadcasted news item is therefore insufficient to fully grasp the context of the fact being reported. In this paper, we propose an approach to retrieve and analyze related documents in order to automatically generate semantic annotations that provide viewers and experts of the domain comprehensive information to fully understand the news content. The approach takes as inputs the publication date and the newscast's title for gathering event-related documents on the Web, bringing more representativeness to the available knowledge. Then named entities detected in the retrieved documents are merged with those found in the newscast subtitles for further disclosing hidden relevant entities that were not explicitly mentioned in the original newscast. A ranking algorithm based on entity frequency, popularity peak analysis, and domain experts' rules sorts the annotations to generate the Semantic Snapshot of the considered Newscast (NSS). We benchmark this method against a gold standard generated by domain experts and assessed via a user survey over five BBC newscasts. Results of the experiments show the robustness of the approach holding an Average Normalized Discounted Cumulative Gain of 66.6%.

Keywords: Semantic Video Annotation, Entity Expansion, Newscasts

1 Introduction

With the emergence of both citizen-based and social media, traditional information television channels have to re-think their production and distribution workflow processes. We live in a globalized world, a vast playing field where events happening are the result of complex interactions between many diverse agents along time. The interpretation of those news is problematic because of two issues. *i)* the *need of background*: viewers probably need to be aware of other facts happened in a different temporal or geographic dimension. *ii)* the *need of completeness*: a single representation of an event is not enough to capture the

whole picture, because it is normally incomplete, it can be biased, or partially wrong. Recent gadgets and applications, such as second screen devices, have recently irrupt as a good way of assisting the viewer in the challenge of becoming aware of that bigger picture of the event. In [4], the authors tracked 260 tablet users, concluding that even though there is a modest uptake and interest in using secondary screens to digitally share opinions, the use of second screen interaction with television content is something the viewer qualitatively appreciate. But there is still a crucial aspect open: the second screen interaction needs to be fed with meaningful details concerning a newscast. The most common strategy to get this information is to enrich the original content with additional data collected from external sources. Today users have access to multiple news portals, different services for commenting and debating on the news, and social media that instantaneously spread news information. However, this results in large amounts of unreliable and repeated information, leaving to the user the burden of processing the large amounts of potentially related data to build an understanding of the event.

Machine driven approaches have tried to alleviate the human difficulties when navigating this huge amount of data, but they struggle both in finding a good set of candidate documents, and filtering them. One strategy reported in the literature for having such a mechanism is to perform named entity extraction over the newscast transcript [10]. However, the set of named entities obtained from such an operation is insufficient and incomplete for expressing the context of a news event [8]. Sometimes 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 some others, entities are simply not mentioned in the transcripts while being relevant for understanding the story. This is also an inherent problem in information retrieval tasks: a single description about the same resource does not necessarily summarize the whole picture. In this paper we automatically retrieve and analyze additional documents from the Web where the same event is also described, in a process called Newscast Named Entity Expansion. By increasing the size of the document set to analyze, 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. This approach is able to produce a ranked list of entities called Newscast Semantic Snapshot (NSS), which includes the initial set of detected entities in subtitles with other event-related entities captured from the seed documents.

The paper is organized as follows: Section 2 presents the related works, Section 3 illustrates the approach depicted in this paper for generating NSS. Section 4 describes in depth the different ranking algorithms used for ordering the list of candidate entities generated in previous steps. In Section 5 we report about the creation of the gold standard used for evaluating the NSS. The experimental settings are described in Section 6. We conclude with Section 7 summarizing the main findings and outlining the future plans.

2 Related Work

The need of a NSS for feeding certain applications is already a concept we have investigated in previous research works and prototypes [12], which have probed the benefit for users of browsing the "surrounding context" of the newscasts. The same concept[3] has been presented to the forum Iberoamerican Biennial of Design (BID).[4] with great feedback from users and experts. Research efforts have underlined the importance in professional journalism of algorithms, data, and social science methods for reporting and storytelling, under the term of computational exploration in journalism [7]. Projects like NEWs have studied how to disambiguate named entity in the news domain by continuously learning while processing news streams [6]. In the domain of Social Networks, named entities are also used for identifying and modeling events and detecting breaking news. In [13], the authors emphasize the importance of spotting news entities in short user generated post in order to obtain a better understanding about what they are talking about. Entities have been used for the video classification when the textual information attached to a video contains temporal references (e.g. subtitles) [9].

In order to build a comprehensive NSS the knowledge expressed in the newscast has to be extended with further entities that are not explicitly mentioned in the video item. Our approach performs an entity expansion process, which allows to collect on the fly event-related documents from the Web. In the literature there are already some approaches relying in similar expansion techniques, even if the final objective is not spotting other event-related entities: instead, they transform the feature space from a from small number of named entities with the same type or category to a more complete named entity set. One of them is Google Sets.[5] Set expansion using the Web is also closely related to the problem of unsupervised relation learning [1], and set-expansion-like techniques have been used to derive features for concept-learning [2], to construct "pseudo-users" for collaborative filtering [3], and to compute similarity between attribute values in autonomous databases [20]. In [17], authors proposed the Set Expander for Any Language (SEAL) approach. SEAL works by automatically finding semi-structured web pages that contain lists of items and the aggregating these list in a way that the most promising items are ranked higher. SEAL is a language-independent system and it has shown good performance against previously published results like the already mentioned Google Sets. By using particular seeds and the top one hundred documents returned by Google, SEAL achieves 93% in average precision in dataset from various languages. The same authors published an improved version of the algorithm [18], increasing the performance by handling unlimited number of supervised seeds. In each iteration, it expands a couple of randomly selected seeds while accumulating statistics from

[3] `https://vimeo.com/119107849`

[4] `http://www.bid-dimad.org`

[5] `http://googlesystem.blogspot.fr/2012/11/google-sets-still-available.html` not longer available since 2014.

one iteration to another. Our approach does not rely on such kind of iterative mechanism, and it focuses on maximizing the quality of the search query for obtaining the most appropriate set of related documents to be analyzed. Another approach in extending set of entities is [15], which combines the power of semantic relations between language terms like synonymy and hyponymy and grammar rules in order to find additional entities in the Web sharing the same category that the ones provides as input. Relying in Google they analyze documents for parsing semi-structured text elements like tables and rank the final candidates using different ranking algorithms like PageRank. Numerous approaches have dealt with a set expansion method using free text rather than semi-structured Web documents; for instance authors in [14] presented a method for automatically selecting trigger words to mark the beginning of a pattern, which is then used for bootstrapping from free text. But still this approach looks for category related entities while in our case the driving force is more the event being shown in the news item. 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 is [11]. It includes a naive document collection strategy, it proposes an entity ranking algorithm based on the appearance of the entities in the collected documents, and it exploits the DBpedia knowledge base as a way to ensure the coherence of the final list of entities. The work presented in this paper improves the referred work in several directions: document retrieval mechanism, semantic annotation, creation of the NSS. [11] is used as one of the baselines in the experimental settings.

3 Newscast Entity Expansion Approach

The approach we use to generate Newscast Semantic Snapshot is composed of the following stages: query formulation, document retrieval, semantic annotation, annotation filtering, and annotation ranking. Fig. 1 depicts the whole expansion process.

Query Formulation Newscast broadcasters offer a certain amount of metadata about the items they publish, which is normally available together with the audiovisual content itself. In this work, we build the query $q = [\mathrm{h}, \mathrm{t}]$, where h is the video heading, and t is the publication date. The query is then used to as input of the retrieval stage.

Document Retrieval The retrieval stage has the intent to collect event-related documents from the open Web. To some extents this process emulates what a viewer, who misses some details about the news he is watching, does: going to the Web, make a search, and get the most of the top ranked documents. Our programmatic approach emulates this human driven task by analyzing a much bigger set of related documents in a drastically smaller amount of time. The stage consists of retrieving documents that report on the same event discussed in the original video as result of the query q. It has an key role in the upcoming semantic annotation stage, since it selects a set of the documents D over which

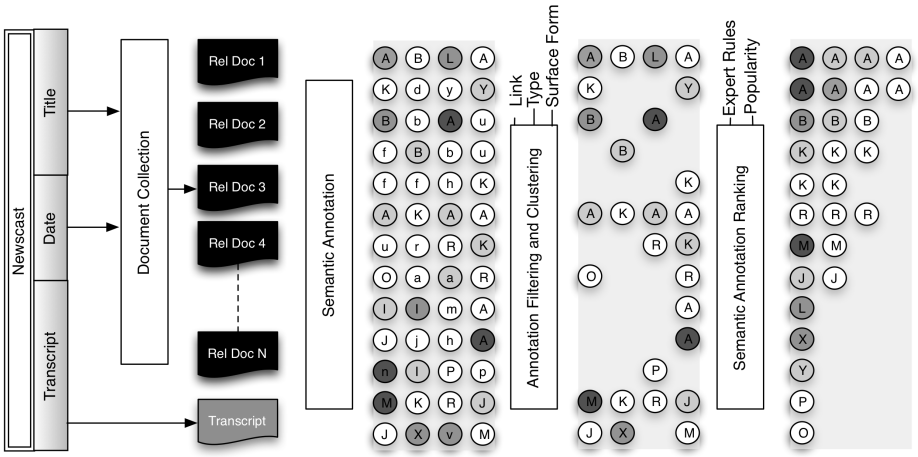

Fig. 1: Schema of Named Entity Expansion Algorithm.

the semantic annotation process is performed. The quality and adequacy of the collected documents sets a theoretical limit on how good the process is done.

Semantic Annotation In this stage we perform a named entity recognition analysis with the objective of reducing the cardinality of the textual content from the set D of documents $\{d_1, ..., d_n, d_{n+1}\}$ where $d_{i=1,..,n}$ defines the i_{th} retrieved document, while d_{n+1} refers to the original newscast transcript. Since most of the retrieved documents are Web pages, HTML tags and other annotations are removed, keeping only the main textual information. 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 $(surface\_form, type, link)$. We perform a union of the obtained bags of named entities resulting in the bag of entities E of the initial query q.

Annotation Filtering and Clustering The Document Retrieval stage expands the content niche of the newscast. At this stage we apply coarse-grained filtering of the annotations E obtained from the previous stage, applying a $f(E_{d_i}) \rightarrow E'_{d_i}$ where $\left|E'_{d_i}\right| < |E_{d_i}|$. The filtering strategy grounds on the findings we obtained in the creation of the gold standard. In fact, when watching a newscast viewers better capture Person-type entities, as well as Organization-type and Location-type entities. The rest of less-specific and wider enclosed entities are more vague to be displayed on a television user interface and potentially less relevant for complementing the seed content. Named entities are then clustered applying a centroid-based clustering operation. As cluster centroid we consider the entity with the most frequent disambiguation *link* that also have the most repeated $surface\_form$. As distance metric for comparing the instances, we applied strict string similarity over the *link*, and in case of mismatch, the Jaro-Winkler string distance [19] over the $surface\_form$. The output of this phase is a list of clusters containing different instances of the same entity.

Semantic Annotation Ranking The bag of named entities E'_{d_i} is further processed to promote the named entities which are highly related to the underlined event. To accomplish such an objective, we propose a ranking strategy based on entity appearance in documents, popularity peak analysis, and domain experts' rules sorts the annotations to generate the Semantic Snapshot of the considered Newscast (NSS).

4 Ranking Strategy

The unordered entity list is ranked to promote the entities that are potentially interesting for the viewer. The strategy we present in this work grounds on the assumption that the entities which appear often in the retrieved documents are important. We propose two different scoring functions for weighting the frequency of the entities. We then consider two orthogonal functions which exploit the entity popularity in the event time window, and the domain experts' rules.

4.1 Frequency-based Function

We first rank the entities according to their absolute frequency within the set of retrieved documents D. Let define the absolute frequency of the entity e_i in the collection of documents D as $f_a(e_i, D)$, we define the scoring function $S_F = \frac{f_a(e_i,D)}{|E|}$, where $|E|$ is the cardinality of all entities spotted across all documents. In Fig. 2 (a) we can observe how entities with lower absolute frequency are placed at the beginning of the distribution and discarded in the final ranking; instead those with high S_F are in the right side of the plot, being then considered to be part of the NSS.

4.2 Gaussian-based Function

The S_F scoring function privileges the entities which appear often. However from the perspective of a television viewer, this is not always the case: while it is true that entities appearing in just a few documents are probably irrelevant and not representative enough to be considered in the final results, entities spread over the whole set of related documents are not necessary the ones the viewers would need to know about. In fact, they often represent entities that have been so present in media before that have become fairly well-known to the viewer. This scoring function is therefore approximated by a Gaussian curve. By characterizing the entities in terms of their Bernoulli appearance rate across all documents $f_{doc}(e_i)$, and applying the Gaussian distribution over those values, we promote entities distributed around the mean $\mu = \frac{|D|}{2}$, being $|D|$ is the cardinality of the retrieved documents (Fig. 2 (b)). In particular the scoring function has been formalized as: $S_G = 1 - \left| \frac{f_{doc}(e_i)}{|D|} - 1 \right|$.

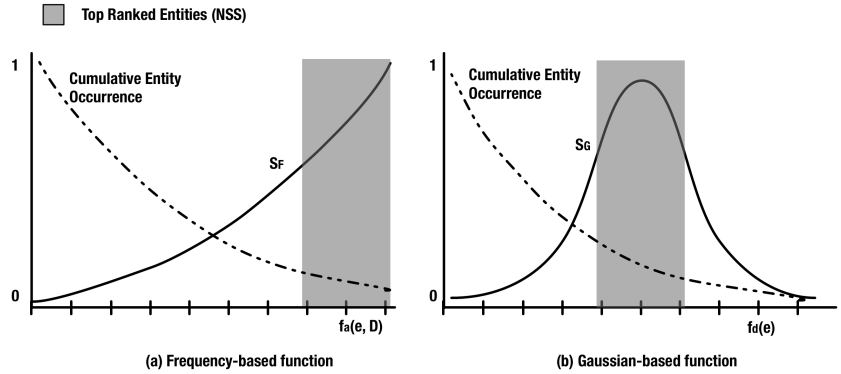

Fig. 2: (a) depicts the Decay function of the entity occurrences in the corpus, and the S_F which underlines the importance of an entity being used several times in the corpus. (b) represents the Gaussian-based function S_G, with the entities highly important over the mean.

4.3 Orthogonal Functions

Popularity Function We propose a weighting function based on a mechanism that detects variations in entity popularity values over a time window (commonly named as popularity peaks) around the date of the studied event. The functions proposed above exploit the frequency of the entities in documents as a factor to measure its importance. However the frequency based approaches fail to explain the phenomenon of certain found entities which are barely mentioned in related documents but suddenly become interesting for viewers. These changes are sometimes unpredictable so the only solution is to rely on external sources that can provide indications about the entity popularity, like Google Trends[6] or Twitter.[7]

The procedure for getting $P_{peak}(e_i)$ is depicted in Fig. 3. Using the label of an entity e_i, we obtain a list of pairs $[t, P]$ where $P \in [0, 100]$ is the popularity score of an entity at the instant of time t. Afterward we create three consecutive and equally long temporal windows around t, the first one w_t containing the date itself, another one just immediately behind w_{t-1} and a last one after the previous two w_{t+1}. In a next step we approximate the area inside the regions by calculating the average of the points contained in them, obtaining $\overline{w-1}$, $\overline{w}$ and $\overline{w+1}$. The slopes of the lines between $\overline{w-1}$ and $\overline{w}$, and $\overline{w}$ and $\overline{w+1}$ give the values m_{up} and m_{down} respectively, which are normalized and combined into a single score for measuring how significant the variation in volume of searches was for that studied entity label. When aggregating those two gradients, we scored m_{up} higher in order to emphasize the irruption of a change, more than the posterior distribution of the search term.

[6] `https://www.google.com/trends`
[7] `https://twitter.com`

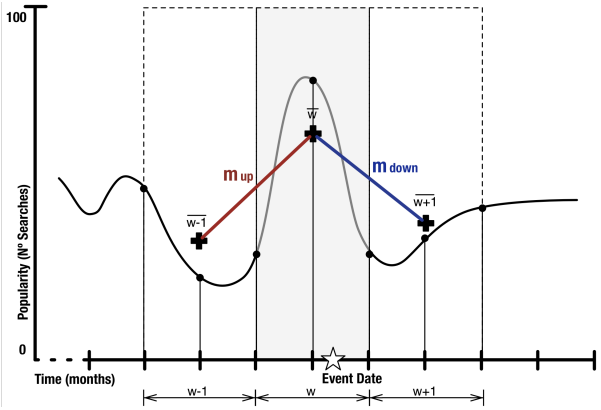

Fig. 3: Popularity diagram of a considered event. On the x-axis is represented the time, and on the y-axis the magnitude of the popularity score. The star indicates when the event occurred. Given the discrete nature of the used platforms, the center of time window can be placed next to the day of the event.

By empirically studying the distribution of the popularity scores of the entities belonging to a newscast, we have observed that it follows a Gaussian curve. This fact will help us to better filter out popularity scores that do not trigger valid conclusions and therefore improving the merging of the ranking produced by the previous functions with the outcome from the popularity peaks detection algorithm.

Expert Rules Function The knowledge of experts in the domain, like journalists or newscast editors can be materialized in the form of rules that correct the scoring output produced by former ranking strategies. The antecedent of those rules is composed by entity entity features such as type, number of documents where the entities appear, or the Web source from where documents have been extracted, while the precedent involves the recalculation of the scoring function according to the following equation: $S_{expert}\,(e) = S_{F-1}\,(e) * Op_{expert}$, being Op_{expert} a factor which models the domain experts' opinions about the entities that match in the antecedent.

5 Gold Standard for Evaluating Newscast Semantic Snapshot

We are interested in evaluating ranking strategies for generating semantic snapshots of newscasts, where each snapshot is characterized by a set of named entities. We narrowed down the selection of named entity types to Person, Organization, and Location, since they can be directly translated in *who*, *what*, *when*, a subset of the fundamental questions in journalism known as the 5Ws. To the

best of our knowledge there is no evaluation dataset suited to this context. The title of the newscasts and the breakdown figures per entity type are shown in Table 1. The dataset is freely available[8].

Newscast Title	Date	Person	Organization	Location	Total
Fugitive Edward Snowden applies for asylum in Russia	2013-07-03	11	7	10	28
Egypt's Morsi Vows to Stay in Power	2013-07-23	4	5	4	17
Fukushima leak causes Japan concern	2013-07-24	7	5	5	13
Rallies in US after Zimmerman Verdict	2013-07-17	9	2	8	19
Royal Baby Prince Named George	2013-07-15	15	1	6	22
Total		46	20	33	99

Table 1: Breakdown entity figures per type and per newscast.

5.1 Newscast Selection

We randomly selected 5 newscasts from the BBC One Minute World News website[9]. Each newscast lasted from 1 to 3 minutes. The selection covered a wide range of subjects specifically: politics, armed conflicts, environmental events, legal disputes, and social news. Subtitles of the videos were not available; therefore, a member of the team manually transcribed the speech in the newscasts.

5.2 Newscast Semantic Annotation

The annotation process involved two human participants: an annotator and a journalist (expert of the domain). No system bias affected the the annotation process, since each annotator performed the task without any help from automatic systems. The output of this stage is a list of entity candidates. The annotators worked in parallel.
The annotator of the domain was asked to detect for each newscast entities from:

subtitle : the newscast subtitle;
image : every time a recognizable person, organization or location was portrayed in the newscast, the entity was added to the list;
image captions : the named entities appearing in such tags, such as nametag overlays, were added to the candidate set;
external documents : the annotator was allowed to use Google Custom Search to look for articles related to the video. The query followed the pattern: title of the newscast, date. The sources were considered: The Guardian, New York Times, and Al Jazeera online (English). The results were filtered of one week time, where the median is represented by the day when event took place.

[8] `https://github.com/jluisred/NewsConceptExpansion/wiki/Golden-Standard-Creation`
[9] `http://www.bbc.com/news/video_and_audio/`

The journalist, with more than 6 years of experience as a writer and editor for important American newspapers and websites, acted as the expert of the domain. He was asked to watch the newscasts and to identify for each the entities either mentioned or not that better serve the objective of showing interesting additional information a final reader. He was completely free to suggest any named entity he wanted.

5.3 Quality control and Ranking

A quality control, performed by another expert of the domain, refined the set of entities coming from the previous stage, eliminating all named entity duplicates and standardizing names. We then conducted a crowdsourcing survey with the objective to gather information about the degree of interestingness of the entities for each newscast. Based on [16] we define interestingness whether an entity is interesting, useful or compelling enough to tear the user away from the main thread of the document. Fifty international subjects participated in this online study. They responded an online call distributed via email and social networks. Their age range was between 25 and 54 years with an average age of 30.3 (standard deviation 7.3 years). 18 participants were female and 32 were male. Most of the participants were highly educated and 48 of them had either a university bachelor degree or a postgraduate degree. The main requisite for participation was that they were interested in the news and followed the news regularly, preferably through means that include newscasts. During the interview participants were asked to choose at least 3 out of 5 videos according to their preferences. Then they were shown each one of the newscasts. Then they were asked to rate whether they would be interested in receiving more information about the named entities in the context of the news video and on a second screen or similar application. All the named entities from the candidate set related to the last seen video were shown in a list with ratio buttons arranged in a similar way to a three-point Likert-scale. The possible answers were “Yes” “Maybe” and “No”.

6 Experimental Settings and Evaluation

In this section we measure the effectiveness of our approach for building the NSS of a newscast against the gold standard shown before. We present the measures considered to carry out the study, we describe the experimental settings, and we conclude with the results.

6.1 Measures

Inspired by similar studies in Web search engines, we have based our evaluation procedure in measures that try to find as many relevant documents as possible, while keeping the premise that the top ranked documents are the most important. In order to summarize the effectiveness of a the different algorithm

across the entire collection of queries considered in the gold standard, we have considered different averaging measures that are listed below:

- Mean precision/recall at rank N. It is probably the most used measure in information retrieval tasks. It is easy to understand and emphasize the top ranked documents. However it does not distinguish between differences in the rankings at positions 1 to p, which may be considered important for some tasks. For example, the two rankings in Figure 4 will be the same when measured using precision at 10.
- Mean average precision at N. Also called MAP, it takes in consideration the order of the relevant items in the top N positions and is an appropriate measure for evaluating the task of finding as many relevant documents as possible, while still reflecting the intuition that the top ranked documents are the most important.
- Average Normalized Discounted Cumulative Gain $MNDCG$ at N. The Normalized Discounted Cumulative Gain is a popular measure for evaluating Web search and related applications [5]. It is based on the assumption that there are different levels of relevance for the documents obtained in results. According to this, the lower the ranked position of a relevant document the less useful it is for the user, since it is less likely to be examined.

As the relevant documents in our gold standard are scored in relevance for the user, we have mainly focused on the last measure since it can provide a more exhaustive judgment about the adequacy of the generated NSS. Concerning the evaluation point N, we have performed an empirical study over the whole set of queries and main ranking functions observing that from $N = 0$ $MNDCG$ decreasingly improves until it reaches a stable behavior from $N = 10$ on.

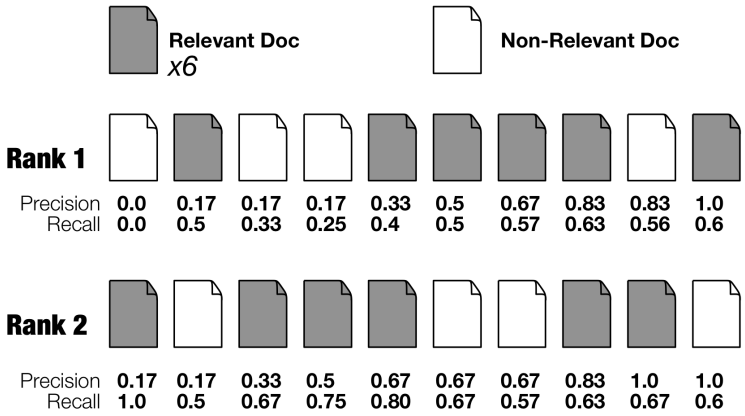

Fig. 4: Inability of P/R for considering the order of the relevant documents: rankings 1 and 2 share the same Precision and Recall at 10 .

6.2 Experimental Settings

Document retrieval We have relied on the Google Custom Search Engine (CSE) API service[10] by launching a query with the parameters specified by $q = [\mathrm{h}, \mathrm{t}]$. Apart of the query itself, the CSE engine considers other parameters that need to be tuned up. First, due to quota restrictions the maximum number of retrieved document is set to 50. But in addition, we have also considered 3 different dimensions that influence the effectiveness in retrieving related documents:

1. Web sites to be crawled. Google allows to specify a list of web domains and subdomains where documents can be retrieved. This reduces the scope of the search task and, depending on the characteristics of the considered sources, influence the nature of the retrieved items: from big online newspapers to user generated content. At the same time, Google allows to prioritize searching over those whitelists while still considering the whole indexed Web. Based on this, in our study we considered five possible values for this parameter:

   Google : search over the whole set of Web pages indexed by Google.
   L1 : A set of 10 internationals English speaking newspapers.[11]
   L2 : A set of 3 international newspapers used in the gold standard creation.
   L1+Google : Prioritize content in Newspaper whitelist but still consider other sites.
   L2+Google : Prioritize content in Lilia's whitelist but still consider other sites.
2. Temporal dimension. This variable allows to filter those documents which are not temporarily close to the day where the newscast was published. Assuming that the news item is fresh enough, this date of publication will also be fairly close to the day the event took place. Taking t as a reference and increasing the window in a certain amount of days d, we end up having $Time_{Window} = [\mathrm{t} - d, \mathrm{t} + d]$ The reason why we expand the original event period is 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 or before. The final $Time_{Window}$ could vary according to 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). In this study we have considered two possible values for it: 2 weeks and one week temporal windows.
3. In addition, Google CSE makes possible to filter result according to the Schema.org types; for our experiments we use the following settings: [NoFilter, Person&Organization]

[10] `https://www.google.com/cse/all`
[11] `http://en.wikipedia.org/wiki/List_of_newspapers_in_the_world_by_circulation`

Semantic Annotation We use [**?**] which applies machine learning classification of the entity type given a rich feature vector composed of a set of linguistic features, the output of a properly trained Conditional Random Fields classifier and the output of a set of off-the-shelf NER extractors supported by the NERD Framework.[12] We used it as an off-the-shelf entity extractor, using the offered classification model trained over the newswire content.

Annotation Filtering and Clustering After some first trials it became evident that there were many non-pure named entities detected in the semantic annotation phase which are not well considered by viewers and experts. We have then applied three different filtering approaches:

F1 : Filter annotations according to their NERD type.[13] In our case, we only keep Person, Location, and Organization.
F2 : It consists in getting ride of entities with confidence score under first quarter of the distribution.
F3 : Intuitively, people seem to be more attracted by proper names than general terms. Those names are normally capitalized. This filter keeps only named entities matching this rule.

By concatenating those filters, we obtain the following combinations: F1, F2, F3, F1_F2, F1_F3, F2_F3, F1_F2_F3). In order to reduce the number of runs, we did a first preselection of filters by setting the rest of steps of the approach to default values and averaging the scores obtained over the different queries. We ended up discovering that 3 of the filters (F1 and F3, and the combination F1_F3) were producing best results in the final MNDCG,

Semantic Annotation Ranking For the current experiment we run both Frequency and Gaussian based functions, together with the orthogonal strategies based on popularity and expert rules. This makes a total of $2*2$ possible ranking configurations that will be considered and reported in result section. Regarding the particular details of the orthogonal functions, we have proceeded as follow:

Popularity We have relied on Google Trends,[14] which estimates how many times a search-term has been used in a given time-window. Since Google Trends gives results with a monthly temporal granularity, we have fixed the duration of such w to 2 months in order to increase the representativity of the samples without compromising too much the validity of the selected values according with the time the event took place. With the aim of being selective enough and keep only those findings backed by strong evidence, we have filtered the entities with peak popularity value higher than $\mu + 2 * \sigma$ which approximately corresponds to a 2.5% of the distribution. Those entities

[12] `http://nerd.eurecom.fr`
[13] `http://nerd.eurecom.fr/ontology`
[14] `https://www.google.com/trends`

will have their former scores combined with the popularity values via the following equation: $S_P(e) = R_{score}(e) + Pop_{peak}(e)^2$.

Expert Rules *i)* Entity type based rules: we have considered three rules to be applied over the three entity types considered in the gold standard. The different indexes per type have been deduced by relying on the average score per entity type computed in the survey $\overline{Sgt}_{entityType}$. Organizations have gotten a higher weight ($Op_{expert} = 0.95$), followed by Persons ($Op_{expert} = 0.74$), and by Locations ($Op_{expert} = 0.48$) that are badly considered and therefore lower ranked in general.
ii) Entity's documents based rules: each entity has to appear at least in two different sources in order to become a candidate. All entities whose document frequency $f_{doc}(e_i)$ is lower than 2 are automatically discarded ($Op_{expert} = 0$).

6.3 Results

Given the different settings for each phase of the approach ($N_{runs_{Collection}} * Runs_{Filtering} * Runs_{Ranking}$), we have a total of $20 * 4 * 4 = 320$ different runs that have been launched and ranked according to $MNDCG_{10}$. In addition we have also executed two baseline approaches for comparing them with the better performing strategies in our approach. More details about them are shown below.

Baseline 1: Former Entity Expansion Implementation As reported in the related work, a previous version of the News Entity Expansion algorithm was already published in [11]. The settings are: Google as source of documents, temporal window of 2 Weeks, no Schema.org selected, no filter strategy applied, and only frequency based ranked function with no orthogonal appliances. Results are reported in Table 2 under the run id *BS1*.

Baseline 2: TFIDF-based Function To compare our absolute frequency and Gaussian based functions with other possible approaches already reported in the literature, we selected the well-known TF-IDF. It measures the importance an entity in a document over a corpus of documents D, penalizing those entities appearing more frequently. The function, in the context of the named entity annotation domain is as follows:

$$tf(e_i, d_j) = 0.5 + \frac{0.5 \times f_a(e_i, D)}{max\{f_a(e'_i, D) : e'_i \in d_j\}}, idf(e_i, d_j) = log\frac{|D|}{\{d_j \in D : e_i \in d_j\}} \quad (1)$$

We computed the average of the TF-IDF for each entity across all analyzed documents, resulting in aggregating the different $tf(e_i, d_j) \times idf(e_i, d_j)$ into a single function $tfidf^*(e_i, D)$ via the function $S_{TFIDF}(e) = \frac{\sum_{j=1}^{n} tf(e,d_j) \times idf(e)}{|D|}$. Results are reported in Table 2 under the run id *BS2*.

Launching the Experiments In Table 2 we present the top 20 runs for our approach in generating NSS, together with some lower configurations at position

Run	Collection			Filtering	Functions			Result			
	Sources	T_{Window}	Schema.org		Freq	Pop	Exp	$MNDCG_{10}$	MAP_{10}	MP_{10}	MR_{10}
Ex0	Google	2W		F1+F3	Freq		✓	0.666	0.71	0.7	0.37
Ex1	Google	2W		F3	Freq		✓	0.661	0.72	0.68	0.36
Ex2	Google	2W		F3	Freq	✓	✓	0.658	0.64	0.6	0.32
Ex3	Google	2W		F3	Freq			0.641	0.72	0.74	0.39
Ex4	L1+Google	2W		F3	Freq		✓	0.636	0.71	0.72	0.37
Ex5	L2+Google	2W		F3	Freq		✓	0.636	0.72	0.7	0.36
Ex6	Google	2W		F1+F3	Freq			0.626	0.73	0.7	0.38
Ex7	L2+Google	2W		F3	Freq			0.626	0.72	0.72	0.37
Ex8	Google	2W		F1+F3	Freq	✓	✓	0.626	0.64	0.56	0.28
Ex9	L2+Google	2W		F1+F3	Freq		✓	0.624	0.71	0.7	0.37
Ex10	Google	2W		F1	Freq		✓	0.624	0.69	0.62	0.32
Ex11	L1+Google	2W		F3	Freq			0.623	0.7	0.72	0.37
Ex12	L2+Google	2W		F3	Freq		✓	0.623	0.68	0.66	0.35
Ex13	L2+Google	2W		F3	Freq	✓	✓	0.623	0.61	0.56	0.3
Ex14	L2+Google	2W		F3	Freq			0.62	0.69	0.74	0.4
Ex15	L1+Google	2W	✓	F1+F3	Freq		✓	0.617	0.69	0.66	0.34
Ex16	L2+Google	2W		F1	Freq		✓	0.616	0.68	0.62	0.32
Ex17	Google	2W	✓	F1+F3	Freq		✓	0.615	0.7	0.64	0.32
Ex18	L1	2W	✓	F3	Freq	✓	✓	0.614	0.65	0.6	0.32
Ex19	L1+Google	2W		F1+F3	Freq			0.613	0.72	0.72	0.38
Ex20	L1+Google	2W		F1+F3	Freq		✓	0.613	0.7	0.66	0.35
...	...	...	...	...	...	...	...	...	...	...	...
Ex78	Google	2W	✓	F1+F3	Gaussian		✓	0.552	0.66	0.66	0.34
Ex80	L2+Google	2W	✓	F1+F3	Gaussian		✓	0.55	0.69	0.7	0.36
Ex82	L1	2W	✓	F3	Gaussian		✓	0.549	0.68	0.64	0.33
...	...	...	...	...	...	...	...	...	...	...	...
BS2	Google	2W			Freq			0.473	0.53	0.42	0.22
...	...	...	...	...	...	...	...	...	...	...	...
BS1	Google	2W			TFIDF			0.063	0.08	0.06	0.03

Table 2: Executed runs and their configuration settings, ranked by $MNDCG_{10}$

78 and following that are worth to be reported and the scores of the baseline strategies. We summarize the main findings of the experimental settings and evaluation as follows:

- Our best approach has obtained a $MNDCG_{10}$ score of 0.662 and a MAP_{10} of 0.71, which are reasonably good in the document retrieval domain.
- Our approach performs much better than BS1 and by far better than BS2. The very low score of this last baseline is explained in the fact that traditional TF-IDF function is designed to measure the relevance of an item referred to the document that contains it and not to the whole collection. In addition, the absence of filters drop drastically the scores.
- Regarding the Document Retrieval step, we see that using Google as source alone of together with other WhiteList gives better results than restricting only to particular whitelist. The biggest T_{Window} of 2 weeks performs better in all cases, while the use of Schema.org does not bring anything back except when is applied over the Gaussian function (see runs 78, 80, 82) where it turns to be an influential factor.

- the best Filter strategy is F3, followed by the combination F1_F3. In conclusion, capitalization is a very powerful tool for making a first candidate list with those entities that a priori users consider more interesting.
- The absolute frequency function performs better than the Gaussian in all top cases.
- The Expert Rules based function improves the final NSS for almost every configuration possible.
- Popularity based function does not seem to improve significantly the results. However, a further manual study of the promoted entities has revealed that in fact, the method is bringing up relevant entities like for example *David Ellsberg*[15] for the query "Fugitive Edward Snowden applies for asylum in Russia". This entity is barely mentioned in some of the collected documents, but *David Ellsberg*'s role in the newscast is quite representative since he published an editorial with high media impact in The Guardian praising the actions of Snowden in revealing top-secret surveillance programs of the NSA.

7 Conclusion

In this paper we have presented an approach for automatically generating Newscast Semantic Snapshots. By following an entity expansion process that retrieves additional event-related documents from the Web, we have been able to enlarge the niche of initial newscast content. The bag of retrieved documents, together with the newscast transcript, is analyzed with the objective of extracting named entities referring to people, organizations, and locations. By increasing the size of the document set, we have increased 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 named entities have been then ranked according to the entity appearance in the sampled collection of documents, popularity of the entity on the Web, and experts' rules. We assessed the entire workflow against a gold standard, which is also proposed in this paper. The evaluation has showed the strength of this approach, holding an $MNDCG_{10}$ score of 0.666, outperforming the two studied baselines.

Future research interests include tailoring the entity ranking functions to particular news categories: sport, politics, regional, international, opinion. We are investigating the role of entity relations in generating of the Newscast Semantic Snapshot: usually entities are linked by tight relations extracted from a knowledge base, or simply from the documents collected, in order to generate a directed graph of entities instead of a list. We also plan to refine the ranking process, applying supervised techniques (Learning to Rank) that tailor the solution on particular domains. In the investigation we have conducted, we have encountered a few missing entities in the gold standard that were not identified because of human limitations when exhaustively covering the whole semantic

[15] `http://en.wikipedia.org/wiki/Daniel_Ellsberg`

picture of the newscast. We then plan to use our approach as a means to suggest relevant entities in the process of the gold standard creation.

Acknowledgments

This work was partially supported by the European Union's 7th Framework Programme via the project LinkedTV (GA 287911).

References

1. M. J. Cafarella, D. Downey, S. Soderland, and O. Etzioni. Knowitnow: Fast, scalable information extraction from the web. In *IN PROCEEDINGS OF THE HUMAN LANGUAGE TECHNOLOGY CONFERENCE (HLT-EMNLP-05*, pages 563–570, 2005.
2. W. W. Cohen. Automatically extracting features for concept learning from the web. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 159–166, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
3. W. W. Cohen and W. Fan. Learning page-independent heuristics for extracting data from web pages. In *In AAAI Spring Symposium on Intelligent Agents in Cyberspace*, 1999.
4. C. Courtois and E. D'heer. Second screen applications and tablet users: Constellation, awareness, experience, and interest. In *Proceedings of the 10th European Conference on Interactive Tv and Video*, EuroiTV '12, pages 153–156, New York, NY, USA, 2012. ACM.
5. W. B. Croft, D. Metzler, and T. Strohman. *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2010.
6. N. Fernndez, J. A. Fisteus, L. Snchez, and G. Lpez. Identityrank: Named entity disambiguation in the news domain. *Expert Systems with Applications*, 39(10):9207 – 9221, 2012.
7. A. Gynnild. Journalism innovation leads to innovation journalism: The impact of computational exploration on changing mindsets. *Journalism*, 15(6):713–730, 2014.
8. M. Henzinger, B.-W. Chang, B. Milch, and S. Brin. Query-free news search. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 1–10, New York, NY, USA, 2003. ACM.
9. 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'13)*, Rio de Janeiro, Brazil, 2013.
10. Y. Li, G. Rizzo, R. Troncy, M. Wald, and G. Wills. Creating enriched youtube media fragments with nerd using timed-text. In *11th International Semantic Web Conference (ISWC2012)*, November 2012.
11. J. L. Redondo Garcia, L. De Vocht, R. Troncy, E. Mannens, and R. Van de Walle. Describing and contextualizing events in tv news show. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 759–764, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

12. J. Redondo-Garca, M. Hildebrand, L. Romero, and R. Troncy. Augmenting tv newscasts via entity expansion. In V. Presutti, E. Blomqvist, R. Troncy, H. Sack, I. Papadakis, and A. Tordai, editors, *The Semantic Web: ESWC 2014 Satellite Events*, Lecture Notes in Computer Science, pages 472–476. Springer International Publishing, 2014.
13. T. Steiner, R. Verborgh, J. Gabarro Vallés, and R. Van de Walle. Adding meaning to social network microposts via multiple named entity disambiguation apis and tracking their data provenance. *INTERNATIONAL JOURNAL OF COMPUTER INFORMATION SYSTEMS AND INDUSTRIAL MANAGEMENT*, 5:69–78, 2013.
14. P. P. Talukdar, T. Brants, M. Liberman, and F. Pereira. A context pattern induction method for named entity extraction. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 141–148, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.
15. M.-V. Tran, T.-T. Nguyen, T.-S. Nguyen, and H.-Q. Le. Automatic named entity set expansion using semantic rules and wrappers for unary relations. In *Asian Language Processing (IALP), 2010 International Conference on*, pages 170–173, Dec 2010.
16. V. von Brzeski, U. Irmak, and R. Kraft. Leveraging context in user-centric entity detection systems. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 691–700, New York, NY, USA, 2007. ACM.
17. R. C. Wang and W. W. Cohen. Language-independent set expansion of named entities using the web. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 342–350, Washington, DC, USA, 2007. IEEE Computer Society.
18. R. C. Wang and W. W. Cohen. Iterative set expansion of named entities using the web. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pages 1091–1096, Washington, DC, USA, 2008. IEEE Computer Society.
19. W. E. Winkler. Overview of record linkage and current research directions. In *Bureau of the Census*, 2006.
20. G. Wolf, H. Khatri, B. Chokshi, J. Fan, Y. Chen, and S. Kambhampati. Query processing over incomplete autonomous databases. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, VLDB '07, pages 651–662. VLDB Endowment, 2007.