eChallenges e-2011 Conference Proceedings
Paul Cunningham and Miriam Cunningham (Eds)
IIMC International Information Management Corporation, 2011
ISBN: 978-1-905824-27-4

A System for Synergistically Structuring News Content from Traditional Media and the Blogosphere

Nikos SARRIS [1*], Gerasimos POTAMIANOS [2*], Jean-Michel RENDERS [3*], Claire GROVER [4*], Eric KARSTENS [5*], Leonidas KALLIPOLITIS [1], Vasilis TOUNTOPOULOS [1], Georgios PETASIS [2], Anastasia KRITHARA [2], Matthias GALLÉ [3], Guillaume JACQUET [3], Beatrice ALEX [4], Richard TOBIN [4], Liliana BOUNEGRU [5]

[1] *Athens Technology Center S.A., 10 Rizariou Street, Halandri, Athens 15233, Greece*
[2] *Inst. of Informatics & Telecommunications, NCSR "Demokritos", Athens 15310, Greece*
[3] *Xerox Research Centre Europe, 6 chemin de Maupertuis, 38240 Meylan, France*
[4] *Inst. for Language, Cognition & Computation, School of Informatics, Edinburgh, EH8 9AB, UK*
[5] *European Journalism Centre, Sonneville-lunet 10, 6221KT Maastricht, The Netherlands*
[*] *Emails: n.sarris@atc.gr, gpotam@iit.demokritos.gr, jean-michel.renders@xrce.xerox.com, grover@inf.ed.ac.uk, karstens@ejc.net*

Abstract: News and social media are emerging as a dominant source of information for numerous applications. However, their vast unstructured content present challenges to efficient extraction of such information. In this paper, we present the SYNC3 system that aims to intelligently structure content from both traditional news media and the blogosphere. To achieve this goal, SYNC3 incorporates innovative algorithms that first model news media content statistically, based on fine clustering of articles into so-called "news events". Such models are then adapted and applied to the blogosphere domain, allowing its content to map to the traditional news domain. Furthermore, appropriate algorithms are employed to extract news event labels and relations between events, in order to efficiently present news content to the system end users.

1. Introduction

News content in the internet, available through both traditional news media portals and the blogosphere, constitutes valuable information to both professionals and casual internet users, who however can be inundated by its vast amount. Clearly, such information could be much more useful if presented and delivered in a well-structured way. Many attempts, taking the form of either research projects or commercial solutions, have been made to provide centralised repositories of such content [1][2][3]. However, to date, there exists no integrated system that structures blog post content across these two broad sources of news information in parallel, capable to meet the requirements of a broad range of end users, such as professional journalists, communication experts, and citizen bloggers. The SYNC3 system[1], presented in this paper, aims to fill this gap, efficiently structuring content from both domains, rendering it accessible, manageable, and re-usable. Following a brief description of the objectives and methodology adopted in SYNC3 (Sections 2 and 3, respectively), the paper focuses in the main SYNC3 algorithmic innovations (Section 4), their integration within a single system (Section 5), and their evaluation (Section 6). Potential business benefits and conclusions are given in Sections 7 and 8, respectively.

[1] This work has received research funding from the European Community Seventh Framework Programme, in the context of the FP7 - 231854 SYNC3 project

2. Objectives

The scope of this paper is to present the innovative ICT solutions that the SYNC3 system introduces in order to efficiently structure content from both traditional news sources and blog posts and present it to relevant stakeholders, ranging from professional journalists and bloggers to communication experts and policy makers. The SYNC3 objective is to take media monitoring and tagging to another level by comparing the latest news from traditional media sources and the blogosphere, enabling users to track their evolution, and to share favourite stories. The paper elaborates on the approach adopted, giving details of the algorithms employed, their integration into a single system, and their evaluation at the component and system level, the latter by target-group users. The paper also highlights the business impact and opportunities from the commercialisation of this system and concludes with potential extensions, which can further leverage its penetration to the target business sector and the respective stakeholders.

3. Methodology

The SYNC3 system is a solution for aggregating news from both traditional news media (i.e. news portals, etc.) and the blogosphere, providing the end users with sophisticated capabilities with respect to content structuring, management, and delivery. The methodology adopted applies the news domain structure derived from well-organised news portals to the less structured blogosphere.

More specifically, SYNC3 automatically builds a news thematology, based on a statistical modelling approach that derives fine clusters of news articles, the so-called "news events". These events are classified into a hierarchy of news topics and themes, based on the IPTC taxonomy [4], and can be further labelled and linked with each other, according to detected temporal, geographical, and causal relations. Subsequently, the system adapts the statistical news event models to the blogosphere domain, allowing the system to automatically find blog posts that comment on these events. Further system components not described in this paper are the "sentiment analysis" module that aims to determine blog post author sentiment towards these events, and a "user interface module" that efficiently presents the extracted information to system users, while also allowing them to update such information individually or collaboratively to meet their needs.

4. Technology Description

SYNC3 consists of three main algorithmic processing chains, which include the analysis of traditional media and news articles (*news processing chain*), the characterisation of news events and their linking to each other (*labelling and relation extraction chain*) and the association of blog posts to news events (*blogs processing chain*). Details are given next.

4.1 The News Processing Chain

The scope of the news processing chain is to analyse news items from professional news sources and categorise them based on the news events that are reflected in them. More specifically, it tries to automatically detect homogeneous groups of documents that report on the same event by means of clustering techniques. Detecting events allows structuring the news sphere in an effective way and, as a consequence, it should allow the user to access the mass of news articles from an event-based viewpoint, rather than a document-based, or a website-based one, as usually done. Through an innovative model for topic and theme categorisation, news events are efficiently clustered into the existing news taxonomy of IPTC. In more detail, the news processing chain consists of four components that are sequentially called, when new content becomes available.

Html cleaning and linguistic pre-processing: This component is in charge of removing irrelevant information in the original html files associated to the news items, as well as parsing the textual content to extract lemmas and named entities (persons, places, organizations) and updating the corresponding dictionaries. So-called "primary sources" (i.e., main news agencies, reporting mostly purely factual information) are cleaned by a finely-tuned rule-based system (based on x-path rules) that, in the same time, is able to extract paragraph information; news items coming from these sources constitute the basis for subsequent clustering algorithms that recognize the news events. For other news sources (so-called "secondary sources") html cleaning is performed by Boilerpipe [5].

Linguistic pre-processing, i.e., tokenization, lemmatisation, named entity recognition/normalisation, and co-reference resolution (both inter- and cross-document) is then applied on the cleaned text, based on the Xerox Incremental Parser technology [6].

Topic/theme categorization: This component probabilistically assigns each news item to one or multiple IPTC codes. It should be noted that, as virtually no data annotated with IPTC codes pre-existed, the building of the categorization models had to follow a non-standard learning strategy, detailed in [7]. The resulting topic/theme categorizer is fully hierarchical, exploiting hierarchical dependencies between IPTC codes.

Event recognition: This component builds event models incrementally, by clustering the "main segments" of news items from primary sources (Note that the main segments consist of the title of a news article, as well as its first paragraphs that, together with the title, form a semantically coherent set). Clustering relies on three non-standard particularities: (a) It tries to be consistent with the clustering results of the previous crawls (similar to the concept of evolutionary clustering [8]): Previous clusters may be updated, while at the same time new clusters – corresponding to potentially new events – are generated. (b) It introduces a forgetting factor that precludes assigning new articles to clusters that have been inactive for several days (an event is assumed to be localised in time). (c) Named entities have their own importance in defining the similarities between segments and clusters: similarities are multi-faceted, in order to capture the fact that articles mentioning the same persons interacting at the same time and in the same location are likely to really define an event. It should be noted that cross-document co-reference resolution is important for this sub-task, as it is often the case that the same entity is expressed with different titles, spellings, extra names, etc. The output of this component is a set of event statistical models, which are then used to classify segments not used for the clustering process (see excerpt extraction, next) or blog posts, after model adaptation. To each event is also associated a weighted set of IPTC codes, which is obtained by aggregating the IPTC code probabilities of the documents that are members of the event.

Excerpt extraction: All segments not used in clustering (i.e., all segments that are not the main segments of articles from primary sources) are then categorized using all active event models, with the possibility of assigning these segments to no events. Note that this actually constitutes a coupled segmentation-categorization problem, since contiguous segments could be more meaningfully assigned to an event than a single (sparse) segment. So, the technology adopted relies both on a dynamic programming technique often used in text segmentation problems [9] and on nearest neighbour classifiers with particular metrics.

4.2 The Labelling and Relation Extraction Chain

This follows the news processing chain, providing additional analysis of the news events. Each news article that is part of an event is fed through a linguistic processing pipeline, including named entity recognition (NER), geo-resolution, and temporal grounding, which are vital for later processing. The news event clusters are then processed by a labelling and a relation extraction component. The former determines document and event-level labels and the latter computes temporal and geographical relations between news events.

The main aim of the news event labelling module is to provide brief descriptions of the news events in terms of their "what" content, thus helping users to find out easily what news events are about [10]. Each news event is given a LABEL (a title-like summary of the news event) and a DESCRIPTION (a one-sentence summary of the event). This information is first computed for every news document (referred to as document summary) and then the most representative document summary for the news event cluster is selected. News titles tend to be appropriate summaries of news items and events. They are coherent phrases or sentences that are understood by users. In order to determine a news event label, variations of title labelling are performed [11], made up of a document-level title detection step followed by event-level title selection step. Given all news document titles, a number of different methods have been adopted to obtain an event LABEL. These include choosing the title of the first published news document, the title of the news item closest to the news event cluster centroid, or the longest/shortest title with the highest term overlap when comparing all titles in the event. For consistency, the DESCRIPTION is extracted by choosing the first sentence following the title that was selected as the LABEL.

The aim of the relation extraction component is to identify "where" and "when" a news event happened, i.e., the news event location and the news event date. This allows users to associate different news events in terms of their geographical and/or temporal relation information. By visualising news events on a map or timeline, users can gain a different perspective on events compared to reading through a flat stream of news. In order to determine the correct news event location of an event, all location entities recognised in the text documents belonging to the news event are first normalised by the Edinburgh Geoparser [12] to a unique GeoNames ID with corresponding latitude and longitude values, as well as additional attributes, such as its population size, its capital or its country name, if appropriate. This step is crucial in order to differentiate between ambiguous place names and ultimately to visualize news events on a map. Given all normalised locations mentioned in the news event, the news event location is then selected using various methods. For example, the most frequent normalised location with the smallest population size mentioned either in the entire event or in all document summaries can be considered as the news event location. News event date extraction follows a preliminary step of temporal grounding [13] of all actual, relative and underspecified temporal expressions extracted from the text in the event. This is done in relation to the document date of the text (i.e. the crawling or publishing timestamp of the feed). Thus each temporal expression is normalised to the correct canonical format (date, month, year, and year attributes) and grounded to a single unique number representation of date. This enables determining the day of the week of temporal expressions, resolving relative dates, and computing temporal precedence. Various methods are employed to select the news event dates amongst all the document dates. Among them, a combination of selecting dates within the text and backing off to the document dates, if the former are not expressed, achieve the highest performance.

4.3 The Blog Post Processing Chain

The blogs processing chain associates blog posts to events, as recognised by the news processing components. A main challenge in this represents the domain shift. Classifying blog posts into events extracted from news items can be easy, if the domain of both blog posts and news items are relatively similar. This can be the case for professional journalists who are also bloggers, as their writing style roughly remains the same when they write news items or blog posts. However, the vast majority of bloggers do not fall into this category, as they typically are individuals expressing personal thoughts, while their writing style may vary significantly from what is observed in the news. In order to associate blog posts from the latter category to the news, an adaptation of the classification model is required, in order to accommodate any possibly new writing styles. This process, known as domain adaptation, must extent a model for handling documents from a different domain (i.e., blog posts), without losing the ability to classify documents from the original domain (i.e., news items replicated in blog posts, or blog posts from journalists).

Blog post processing commences with html cleaning and linguistic pre-processing, similar to the ones employed by the news processing chain. The title and text are extracted from each blog post using Boilerpipe [5]. Then, posts are segmented, and named entities, as extracted by the news processing components, are located. The resulting posts are subsequently used as a corpus that drives model adaptation in an unsupervised fashion.

Since news processing creates a statistical event model, the domain adaptation process can be divided into two tasks. The first aims in expanding the feature space, so as to include new features from the blogosphere domain, while the second concentrates on how weights can be adapted, so as to maximise classification performance on the union of the two domains. For the task of feature space expansion, an approach based on text relatedness has been developed, which locates features from the blogs domain that are "related" to existing features from the news domain, according to a text relatedness metric. This metric is based on WordNet synonimity [14], and two text segments are related if the intersection of their synonyms is not empty. New features, extracted from the blogs domain, are discarded if they are related to more than one original feature from the news domain, and their weights are initialised by inheriting the weights of their related, original feature from the news domain. The second task of model adaptation aims to optimise the weights of the events model, again in an unsupervised manner. Weight adaptation is guided by the differences between the two domains, through the Kullback-Leibler Importance Estimation Procedure (KLIEP) algorithm, which tries to estimate the ratio of two density functions without calculating density

estimations [15]. Once adapted statistical events models are acquired, the k-nearest neighbour algorithm (kNN) is employed to classify blog posts, using k=1 and cosine similarity as a distance metric.

5. Developments

SYNC3 system development has now reached its second prototyping phase, in which the envisaged functionalities for the three processing chains have been implemented and are being improved. Logically, the system has been designed on a layered architecture, as shown in Figure 1, which serves the general principles of modern enterprise systems through a service-oriented architecture.

In more detail, the *service access component* acts as the orchestrating software module of this architecture and integrates the individual process chains analysed in Section 4. This component has been developed so as to link the components laid on the different layers. The *data access layer* is based on object-relational mapping (ORM) and contains all the data access objects that link the core application with the available repositories. The *business layer* acts as the core of the system with all business logic implemented at this level. Crawling of the news and blog sources, use of the metadata API, and complex database manipulation take place in this layer. All these actions make use of the data access layer components for the low level connection between the real data items. Finally, the *service layer* involves all the resources exposed via the http protocol by the server of the system. These resources consist of the RESTful web services that expose the collected and processed data from the multimedia and the metadata repositories. The *presentation layer* makes use of the available services and the query API for the metadata repository in order to retrieve all the information required to provide a complete answer to user requests. All these layers are harmonically connected to provide the functionalities of the SYNC3 system through the dataflow diagram, which is depicted in Figure 2. For more information, please refer to [16].

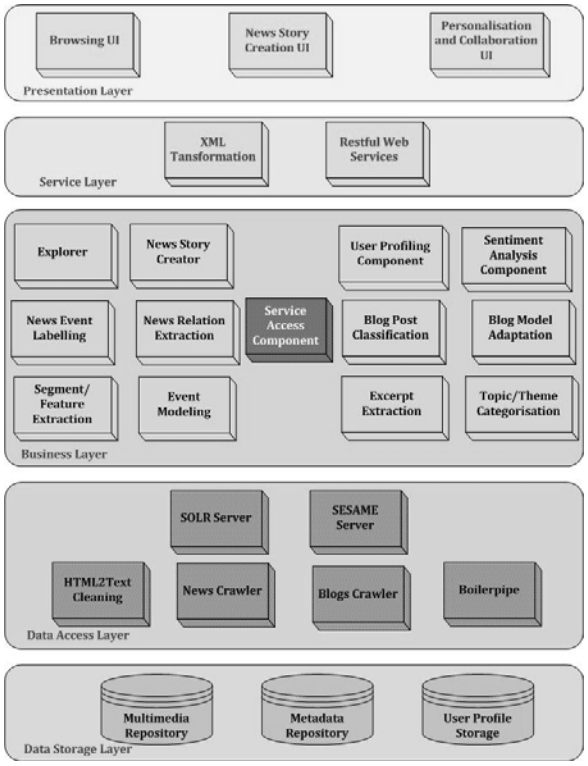

Figure 1: The Architecture of the SYNC3 System

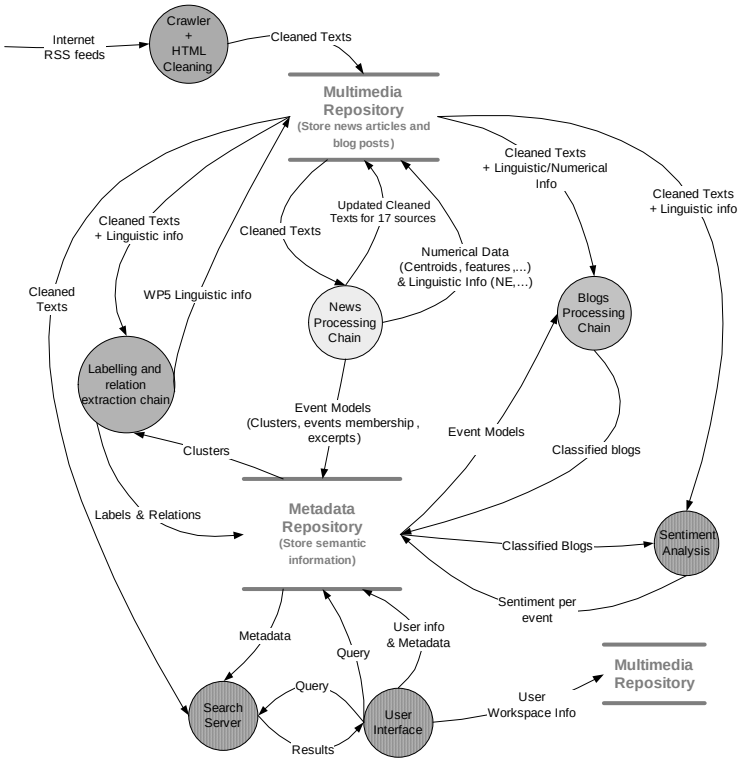

Figure 2: The Dataflow Diagram of the SYNC3 System

6. Results

The functionalities and underlying technologies of the integrated SYNC3 system have undergone a two level evaluation for assessing the technical maturity of the developments and the target

stakeholder perception and acceptance of the functionalities offered[2]. As illustrated in Figure 3, the prototype offers a free-text area to allow end users to submit simple keyword-based queries (1). The result of the free-text search is the list of news events, which have been first identified from the news items corpora and match the end user query (2). For each event, the end user can view the available metadata information, which has been associated with the specific event (3). Furthermore, each of the news events has been linked to related news articles and blog posts (4). The sentiments associated to the specific event are also visualised (5). The users can finally filter the results by selecting one or more of the recognised Named Entities listed in the left column (6).

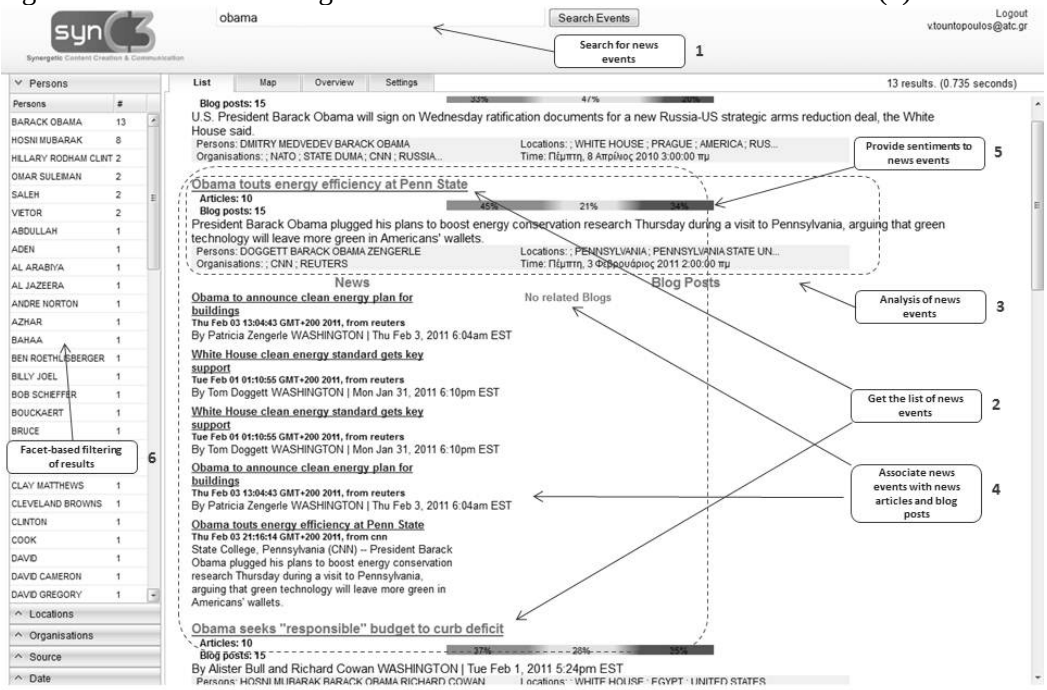

Figure 3: The SYNC3 User Interface

6.1 User Evaluation Results

A first version of the SYNC3 system has undergone usability and functionality testing in one-to-one sessions with 21 test participants from key end-user target groups, namely media analysts, journalists, editors, bloggers, and media consumers by using the "think aloud" method. To collect feedback in a quantifiable manner, a questionnaire was distributed to the test participants at the end of the testing sessions.

The concept and intention exposed by the system developments have been approved by the test participants. Particularly, the capacity to enable better understanding of the dynamics between traditional and social media by linking together news articles with blogs that relate to them was well-received. Observation of uninitiated users interacting with the system, through the provided intrusive interface, yielded the overall impression that they quickly grasped its purpose and main functions. Users appreciated that the tool was clean and clear visually. Over half of the respondents rated favourably the speed of the system. Usability aspects have generally received positive ratings as they have been answered by using the points 9 to 5 pertaining to the positive side of options on the 0 to 9 response scale, meaning that there were no major frustrations regarding the interaction with the system and the interface layout.

In terms of functionality, the results of the initial evaluation were critical, which reveals the need for further improvements, in order for the system to be commercially accepted. However, two thirds of the respondents rated favourably the accuracy of the event labels in describing the news events. 30% of the respondents considered the generated results to be sufficiently relevant to their queries, while 60% of the respondents considered the relevance of the generated results to require further improvement. All in all, the rating of the system as "needing improvement" suggests that

[2] All results should be considered preliminary as system development is still in progress.

the concept behind the SYNC3 system has been met with the approval of the test participants, which was one of the main objectives of the first evaluation, and reflects the naturally "raw" status of a system in its first prototypical stage.

6.2 Technical Evaluation Results

Following the methodological approach identified above, this section presents the results from the technical evaluation of the individual components, which are integrated into the SYNC3 system. In order to do so, a set of manually annotated data has been created, consisting of news articles from primary news sources and blog posts from creditable sources, randomly crawled from the internet and annotated based on the project needs.

With respect to the news processing chain, the event recognizer (clustering), the excerpt extraction, and the topic/theme (IPTC codes) categorizer have been evaluated. With respect to the clustering algorithm, a total set of 185 news articles have been used, forming 44 news events. To compare the clustering results with the manually annotated events, the following standard approach was chosen: each (gold-) event e is assigned to the cluster $\sigma(e)$ that maximizes the corresponding micro F_1 measure between e and $\sigma(e)$. The final results are shown on Table 1, using standard performance metrics to compare two partitions. The total number of identified clusters was 43. A similar approach has been adopted for evaluation of excerpt extraction. The mapping σ between gold-events and clusters is fixed right after clustering as before. After each document undergoes excerpt extraction, each article is labeled with those clusters assigned to one of its excerpts (if any). During a real-time analysis of the news media, some of the articles may not refer to any of the identified clusters, because they are alternative or local news not reported in the primary sources. To simulate this "open world" (as opposed to a "closed world") situation, 274 articles reporting news four months later were added to the annotated set. As Table 1 shows, the presented algorithm is robust to this kind of noise.

Table 1: Performance of the methods proposed to recognize manually annotated events and to correctly extract event-oriented excerpts from news articles.

	Clustering	**Excerpt Extraction**	
		closed	**open**
micro P	0.8696	0.6926	0.6873
micro R	0.9730	0.8911	0.8957
micro F1	0.9184	0.7794	0.7778
macro P	0.8750	0.6950	0.6846
macro R	0.9676	0.8754	0.8699
macro F1	0.8903	0.7413	0.7321

Finally, as far as the topic/theme categorizers are concerned, a separate, broader collection of 1100 news articles (from two main news agencies in Europe) has been used, which has been independently labelled by journalists using the IPTC taxonomy. It should be noted that model training did not use this labelled data (neither any IPTC-coded news articles). The hierarchical-F1 measure for this collection reaches 67%, which is quite satisfying given that the classifier has the choice between more than 1100 categories.

An initial evaluation on the news event labelling has been performed by comparing the system label against the manually extracted gold label. This is done both automatically using the Rouge-1 metric [17] as well as manually. For a preliminary test set of 36 news events, the automatic evaluation results in an average F-score of 0.36 for the currently best performing algorithm of selecting the shortest most representative title of the news articles in the event. In the manual evaluation only 2 system labels were deemed completely incorrect, all others were classed as correct or almost correct (7), acceptable (19) and partially acceptable (8). The geographical relation information is extracted for 47 news events with an accuracy of 51.1% for strict GeoNames ID matching, 59.6% for location string matching or 83.0% for a more lax location string or country matching. The temporal relation information is extracted with an accuracy of 83.0%.

Finally, regarding the blog post processing chain, the two tasks of domain adaptation, namely feature space expansion and weight re-estimation, have been evaluated separately. The task of feature space expansion succeeded in expanding an original feature space of 15686 features from

the news domain, with 3018 features from the blogs domain, through text relatedness. The classification accuracy of the adapted events model was measured at 88.79% when classifying blog posts. The task of weight re-estimation was applied on the original feature space, as extracted from the news domain with the help of the news processing components. The classification accuracy of the re-weighted events model was 91.95% when classifying blog posts.

7. Business Benefits

The SYNC3 system can be commercially exploited to analyse news and blog sources and provide a comprehensive roadmap to news and story creation, relevant opinions, and sentiments expressed for news events in the blogosphere. Through this system, news and media organisations, as well as other relevant stakeholders, such as research institutions dealing with media analysis, could potentially monitor the social media environment and participate in fostering opinions in a local, national, and international level. The system significantly improves the way that the vast blog content can be accessed putting them into context through links with the corresponding events documented in the official news sources, thus lifting the barrier in the way towards a new era of effective communication among citizens and synergetic formation of public opinion, an idea that has long been evangelized by media and internet experts alike. The SYNC3 system, as an enabling system, can be the last piece in the puzzle needed for materializing the concept of collaborative structuring of the public opinion.

8. Conclusions

This paper presented the SYNC3 system, which has been developed to combine news content from both traditional news sources and the blogosphere. It analysed the technological advances and presented initial results, through evaluation on a manually annotated dataset. These initial results provide an encouraging step towards integrating state-of-the-art algorithms with a functioning system addressing fundamental business needs. The technical-level results show that the algorithms can work well, given the appropriate training dataset, which refers to a wide range of domains.

References

[1] *Europe Media Monitor (EMM) News Explorer* [Online] http://emm.newsexplorer.eu
[2] *Silobreaker Premium* [Online] http://info.silobreaker.com
[3] *Thoora Service* [Online] http://thoora.com
[4] Metadata Taxonomies for the News Industry. *International Press Telecommunications Council (IPTC)* [Online] http://www.iptc.org
[5] C. Kholschutter, P. Fankhauser, and W. Nejdi, "Boilerplate detection using shallow text features". *Proc. WSDM*, 2010.
[6] S. Ait, J.P. Chanod, and C. Roux, "Robustness beyond shallowness: Incremental dependency parsing". *NLE Journal*, 2002.
[7] V. Ha-Thuc and J.M. Renders, "Large-scale hierarchical text classification without labelled data", *Proc. WSDM*, 2011.
[8] D. Chakrabarti, R. Kumar, and A. Tomkins, "Evolutionary clustering". *Proc. KDD*, 2006.
[9] P. Fragkou, V. Petridis, and A. Kehagias, "A dynamic programming algorithm for linear text segmentation". *Journal of Intelligent Information Systems*, 23(2), 2004
[10] B. Alex and C. Grover. "Labelling and spatio-temporal grounding of news events". *Proc. NAACL* 2010.
[11] C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
[12] R. Tobin, C. Grover, K. Byrne, J. Reid, and Jo Walsh. "Evaluation of georeferencing". *Proc. GIR*, 2010.
[13] C. Grover, R. Tobin, B. Alex, and K. Byrne. "Edinburgh-LTG: TempEval-2 system description". *Proc. SemEval*, 2010.
[14] G.A. Miller, "WordNet: a lexical database for English". *Comm. ACM*, 38(11), 1995.
[15] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, "Direct importance estimation for covariate shift adaptation". *Ann. Inst. Statistical Mathematics*, 60(4), 2008.
[16] *The SYNC3 Project* [Online] http://www.sync3.eu
[17] C.-Y. Lin. "ROUGE: a package for automatic evaluation of summaries." *Proc. WAS*, 2004.