Report on the 2nd International Workshop on Recent Trends in News Information Retrieval (NewsIR'18)

Dyaa Albakour[1]
Signal Media, London, UK

David Corney
Factmata, London, UK

Julio Gonzalo
UNED, Madrid, Spain

Miguel Martinez[1]
Signal Media, London, UK

Barbara Poblete
University of Chile, Santiago, Chile

Andreas Vlachos
University of Sheffield, Sheffield, UK

[1]{*firstname.lastname*}*@signalmedia.co*

Abstract

The news industry has undergone a revolution in the past decade, with substantial changes continuing to this day. News consumption habits are changing due to the increase in the volume of news and the variety of sources. Readers need new mechanisms to cope with this vast volume of information in order to not only find a signal in the noise, but also to understand what is happening in the world given the multiple points of view describing events. These challenges in journalism relate to Information Retrieval (IR) and Natural Language Processing (NLP) fields such as: verification of a source's reliability; the integration of news with other sources of information; real-time processing of both news content and social streams; de-duplication of stories; and entity detection and disambiguation. Although IR and NLP have been applied to news for decades, the changing nature of the space requires fresh approaches and a closer collaboration with our colleagues from the journalism environment. Following the success of the previous version of the workshop (NewsIR'16), the goal of this workshop, held in conjunction with ECIR 2018, is to continue to stimulate such discussion between the communities and to share interesting approaches to solve real user problems. A total number of 19 submissions were received and reviewed, of which 12 were accepted for presentation. In addition to that, we had over 30 registered participants in the workshop who were pleased to attend the two keynote talks given by well-known experts in the field - Edgar Meij (from industry) and Peter Tolmie (from academia) and oral and poster presentations from the accepted papers. The workshop also included a breakout session to discuss ideas for a future data challenge in news IR and closed with a focused panel discussion to reflect on the day. In summary, several ideas were presented in the workshop on solving complex information needs in the news domain. In addition, the workshop concluded with suggestions of important challenges and shared tasks to work on as a community for News IR.

1 Introduction

The news ecosystem is one of the parts of our society that has changed significantly in the last decade. It has changed from a relatively small community with a limited range of reputable and accountable newspapers and broadcasters that provided their perspective on world affairs. Currently, the situation could not be more different: virtually anyone can act as a journalist and spread their point of view to millions of people across the globe using social networks or by having an influential blog. The number of sources of news has soared and now we have millions of data sources producing content. This abundance of information has created several challenges that cannot be addressed without automatic or semi-automatic solutions. Examples of these challenges includes, but are not limited to, the capability of processing hundreds of millions of text documents in effectively real-time; identifying bias in news reporting, measuring credibility of news sources and automatically discovering if they are presenting inaccurate or misleading information; aggregating and summarising opinions from multiple news articles reporting on the same event.

From the journalism perspective, the implications of the above are huge. Also, one major shift that is changing the way journalism works is the use of automatic systems either for writing complete stories (e.g., share price fluctuations) or for supporting journalists and editors reviewing their pieces or collating material so they have all the relevant data. This is a critical factor in a world where authors sometimes have just a few minutes to accurately report on ongoing events, while still making a piece worth reading for the consumers.

We strongly believe that members of the IR/NLP community and professional journalists can collaborate effectively to improve this situation. Therefore, and following success and recommendations from the NewsIR'16 workshop, we organised this workshop in conjunction with ECIR 2018. The aim is to stimulate discussion around the current challenges in the journalism and news processing environment, such that we can combine the expertise of two communities, namely the Information Retrieval community and the Journalism community. In the call of papers, we encouraged submissions on multiple IR tasks that can help solve problems for journalists in this domain.

2 Workshop Programme

The Workshop programme included two keynote talks representing both academia and industry, paper presentations and a poster session, finishing with two breakout groups followed by a panel. The call for papers attracted a total of 19 submissions which included both long technical papers and short position and demo papers. It should be noted that submissions were diverse covering groups from different continents (America, Europe, and Asia) and different research communities. Each submitted paper was reviewed by at least three members of the programme committee and a meta-review was provided by one of the organisers. This was followed by a discussion period among the workshop organisers as a result of which decisions about acceptance/rejection were made and a total of 12 papers were accepted. The full workshop proceedings were published in the CEUR Workshop Proceedings[1].

The one-day workshop was very well attended with over 30 registered participants and was structured as four sessions: two in the morning and two in the afternoon. In the first session, the workshop was kicked off with an introduction from Dyaa Albakour (Signal Media)

[1] http://ceur-ws.org/Vol-2079/

providing a background and a summary of the overall objectives for holding the workshop. Following this, the first keynote talk was presented by an industry representative, Edgar Meij from Bloomberg. After the keynote, authors of all papers presented short talks. In addition, the authors had the opportunity to discuss their work in detail with attendees during a poster session within the second session of the morning. The third session of the day started with an academic keynote by Peter Tolmie from Universität Siegen. For the last session, Dyaa Albakour (Signal Media) presented the topic of the breakout group discussion where two groups were asked to discuss challenges and requirements for a modern data challenge track in news IR. A representative from each group presented their findings to the wider audience. Finally, the workshop was closed with a panel discussion managed by Miguel Martinez (Signal Media). The panel reflected on the day and the various research challenges pointed out during the workshop.

In the following, we detail the workshop's activities outlined above. To give the reader a flavour of the workshop's presentations and discussions. In particular, to structure the discussion, we identify three main themes: *Credibility, Bias and Reputation*, *Media Monitoring* and *News Ranking & Recommendation*.

2.1 Theme 1: Credibility, Bias and Reputation

Edgar Meij kicked off the discussion of this topic with a keynote titled "AI & Automated News: Implications on Trust, Bias, and Credibility". Edgar started by introducing the concept of robotic journalism where news reports can be automatically generated by algorithms. According to Edgar: "everyone has seen a generated news article". Examples of those include stock market reports, results of sport events, natural disasters and weather forecasts. These systems are usually based on rules and templates for language generation. Another form is generation of stories based on images. The main benefits include the almost instantaneous generation of news based on structured data (e.g., stocks value) in addition to freeing up time of journalists to work on more complex tasks rather than routine reporting. Also, this opens the doors to a situation where the news are hyper-personalised for each individual, telling a slightly different story, or at least using a different style. However, robotic journalism has few limitations. First, it may raise moral and ethical questions. For instance, what if the system incorrectly writes that the wrong city is on the way of a tsunami? This could cause panic and chaos. Also, robotic journalism currently lacks opinions and qualitative reasoning which is a potential direction for the community to work on.

After the keynote, a number of paper presentations tackled various aspects of news credibility and bias. An interesting idea was presented by [1] to introduce an "information nutrition label". Inspired from "typical" dimensions used to describe the nutrition values of food (e.g., carbohydrates, sugar, proteins), quality dimensions can be assigned to a news article to measure multiple aspects such as the readability, verbosity, and virality of the article and the accuracy of the source. They also proposed an iconography to display them. To this end, one way to estimate the credibility of a source is to use the accuracy of claims made by them as presented by [2]. For instance, if a newspaper constantly predicts specific situations to happen (e.g., a member of the White House administration will be fired within a month), while never really being materialised, their credibility should be low. In the same line of research, the work presented by [3] aims to estimate the polarity of certain sources based on historical stance analysis on the topic. Finally, another paper looked at the ancillary copyright for snippet generation [4]. The authors discussed the ongoing legal situation in

Germany which is trying to regulate how search engines and other websites re-use contents generated by others. To address this, the authors suggest an "technological remedy" by synthesizing true original snippets without text reuse.

2.2 Theme 2: Media Monitoring and Social Media

One aspect that was clear in the workshop is that users are no longer limited to major media outlets, nor only traditional newspapers for that matter. Social media, blogs and community journalism are becoming increasingly important for many cases related to media monitoring. This requires the collaboration of multiple communities both in the academic side and the journalistic one, and one of the critical factors is to involve journalist experts when we design information systems for the space. Indeed the keynote by Peter Tolmie from Universtät Siegen entitled "Every tool is better than nothing?: The use of dashboards in journalistic work" discussed some work in this direction. Part of the talk described the outcomes of the EU Pheme project on assessing the veracity of claims online. Pheme developed online tools and dashboards that can be used by journalists to explore and visualise rumors in social media. To this end, a number of papers presented tools developed for journalists to explore news. The authors of [5] presented Cross-Reading News, an online tool that can aggregate information from multiple news sources and provide a cohesive summary for a journalist interested in a certain topic or an entity. The authors were brave enough to show a live demo to the audience using the Signal Media one million article collection as their data source . Furthermore, the work presented in [6] took a step further where a video summary is generated for a news event. This is done by identifying events with a doc2vec-based clustering approach and then retrieving images in the web that best match these clusters. The system presented in [7] show how different entities interact in the news via a graph visualisation [2].

Finally, other aspects of media monitoring discussed in the workshop include entity detection in languages with limited resources (the Teglu language) [8], and comparing information consumption between social and mainstream media [9].

2.3 Theme 3: News Ranking and Recommendation

In the light of the discussed topics above, it is clear that there are challenges and opportunities for a variety of ranking and recommendation tasks in the news domain. Deep learning approaches were proposed for such tasks. The work presented in [10] proposed a neural network architecture for news filtering to address limitations of collaborative filtering and content-based filtering for recommendation. Such approaches rely on word embedding representations that can capture latent semantics of the words. The work presented in [11] argued that for news retrieval tasks, such embeddings can become out-of-date with new events and new vocabulary appearing. For example, Las Vegas is now associated with gun control, but a year ago it was not. They show empirically that an updated word embedding model perform better than an older model or a static model for a news retrieval task.

Finally, the last talk of the day presented the TREC 2018 News Track, which proposes two tasks [12]. The first task is background linking where the goal is to retrieve related articles that can provide context to a given article. The second task is ranking the entities in an given article according to their importance. The data collection for this track will be 5-years worth of articles from the Washnigton Post.

[2] A running demo is available here `http://newsir-demo.ifi.uni-heidelberg.de/`

2.4 Breakout Groups

For the breakout groups, to initiate the discussion, we raised hypothetical questions to the participants. The questions were:

- If we organise this workshop next year, and we want to introduce a shared task (challenge) to work on, what will the task(s) be?
- What datasets do we need to evaluate it?

Answering these question will help us in understanding the priorities of the community around news IR in addition to building datasets and evaluation frameworks to foster research in this area. Participants formed three groups and each group discussed the questions and reported their findings to the workshop. We summarise their findings in the following:

- The groups emphasized the importance of working on a task relevant to professional users such as journalists.
- The importance of giving users a diverse view of all the opinions out there was discussed in multiple talks during the day. Therefore, one group proposed to work on a number of tasks to solve this problem. This includes identifying the veracity and the neutrality of news sources. In addition, a summarisation task was proposed where the objective is to provide a diverse set of opinions in the news for a certain topic. Furthermore, it was suggested that this task should be first broken down into simpler tasks. For example, instead of finding and clustering all opinions out there, we can start by understanding whether two articles share the same opinion. It was also noted for this to work, the datasets used should cover different media sources and media types and multiple languages, such that minorities are well represented.

2.5 Panel

For the panel, we invited our two keynote speakers, Edgar Meij and Peter Tolmie to be joined by two co-organisers Barbara Poblete and David Corney. This panel represents the industry and the academic world of Information Retrieval, as well as a perspective from the journalistic environment. The questions raised to the panelists stemmed from the various talks and the discussions during the day.

One of the first discussion points for the panel was the apparent "convergence" of different types of content and how the differences between traditional media, blogs, social media, citizen journalism and other sources of news are becoming less clear. This heterogeneity of types of knowledge should be reflected in the way we process this information in our field. The panelists agreed on this point but also reminded us that there are still pivotal differences between some of those types of news. For instance, traditional newspapers are still held to a much higher standard than blogs in terms of verification, fact checking and multiple sources for a given story.

The mention of credibility and trust was a great entry point to one of the trendiest topics in technology (and society) in recent years. The panel was asked about the impact of disinformation and misinformation (sometimes vaguely referred to as Fake News) and how we should respond as a community. This was an interesting debate with some of the panelists being sure that it will change our field significantly, while others believe it is just a recent spike of interest in a topic that is as old as news itself. Whether the problem of disinformation (or its perception) is growing or decreasing, all of the panelist agreed that we,

as a community, could and should help. However, we should be careful to not become censors ourselves. The panel preferred to follow an approach where we could inform the public about specific characteristics of the news (e.g. biased, inaccurate, etc.) and then, they could make their own mind. This resonates quite close to one of the papers presented at the event that proposed an "Information Nutrition Label" [1].

Another important point of discussion was the integration of the journalistic and the Information Retrieval communities. One of the main conclusions was that we must integrate any of our solutions on their workflow in order to be effectively used. This implies that we should involve journalists, media experts and editors early on in our research ideas looking for validation about the problems we are solving and the systems we are designing. Not doing this was identified as one of the main causes of project failures in the past.

As a summary of the panel discussion, the News IR space is not only still relevant, but it has several newer problems that make it a very interesting and attractive space to work on.

3 Conclusions

Overall, the workshop was a great success according to various indicators: the large number of submissions, the quality of the papers, the diversity of topics, the large number of attendees and the high level of interactions in the various discussions throughout the workshop. This is an indication of the continuing interest of the community in news IR. Building on this, we aim to follow it up by another edition in 2019, potentially in a bigger conference such as SIGIR 2019, or as part of an evaluation framework such as CLEF.

Acknowledgments

We thank all the PC members for their constructive reviews. A detailed list of their names and affiliations can be found on the workshop's website[3].

References

[1] Tim Gollub, Martin Potthast, and Benno Stein. Shaping the information nutrition label. In *NewsIR@ECIR*, volume 2079 of *CEUR Workshop Proceedings*, pages 9–11. CEUR-WS.org, 2018.

[2] Navya Yarrabelly and Kamalakar Karlapalem. Estimating credibility of news authors from their WIKI validated predictions. In *NewsIR@ECIR*, volume 2079 of *CEUR Workshop Proceedings*, pages 12–17. CEUR-WS.org, 2018.

[3] Masaharu Yoshioka, Myungha Jang, James Allan, and Noriko Kando. Visualizing polarity-based stances of news websites. In *NewsIR@ECIR*, volume 2079 of *CEUR Workshop Proceedings*, pages 6–8. CEUR-WS.org, 2018.

[4] Martin Potthast, Wei-Fan Chen, Matthias Hagen, and Benno Stein. A plan for ancillary copyright: Original snippets. In *NewsIR@ECIR*, volume 2079 of *CEUR Workshop Proceedings*, pages 3–5. CEUR-WS.org, 2018.

[3] http://research.signalmedia.co/newsir18/

[5] Shahbaz Syed, Tim Gollub, Marcel Gohsen, Nikolay Kolyada, Benno Stein, and Matthias Hagen. Cross-reading news. In *NewsIR@ECIR*, volume 2079 of *CEUR Workshop Proceedings*, pages 24–26. CEUR-WS.org, 2018.

[6] Alberto Barrón-Cedeño, Giovanni Da San Martino, Yifan Zhang, Ahmed M. Ali, and Fahim Dalvi. Qlusty: Quick and dirty generation of event videos from written media coverage. In *NewsIR@ECIR*, volume 2079 of *CEUR Workshop Proceedings*, pages 27–32. CEUR-WS.org, 2018.

[7] Erich Schubert, Andreas Spitz, and Michael Gertz. Exploring significant interactions in live news. In *NewsIR@ECIR*, volume 2079 of *CEUR Workshop Proceedings*, pages 39–44. CEUR-WS.org, 2018.

[8] SaiKiranmai Gorla, Sriharshitha Velivelli, N. L. Bhanu Murthy, and Aruna Malapati. Named entity recognition for telugu news articles using naïve bayes classifier. In *NewsIR@ECIR*, volume 2079 of *CEUR Workshop Proceedings*, pages 33–38. CEUR-WS.org, 2018.

[9] José Luís Devezas and Sérgio Nunes. Social media and information consumption diversity. In *NewsIR@ECIR*, volume 2079 of *CEUR Workshop Proceedings*, pages 18–23. CEUR-WS.org, 2018.

[10] Dhruv Khattar, Vaibhav Kumar, Manish Gupta, and Vasudeva Varma. Neural content-collaborative filtering for news recommendation. In *NewsIR@ECIR*, volume 2079 of *CEUR Workshop Proceedings*, pages 45–50. CEUR-WS.org, 2018.

[11] Taewon Yoon, Sung-Hyon Myaeng, Hyun-Wook Woo, Seung-Wook Lee, and Sang-Bum Kim. On temporally sensitive word embeddings for news information retrieval. In *NewsIR@ECIR*, volume 2079 of *CEUR Workshop Proceedings*, pages 51–56. CEUR-WS.org, 2018.

[12] Shudong Huang, Ian Soboroff, and Donna Harman. TREC 2018 news track. In *NewsIR@ECIR*, volume 2079 of *CEUR Workshop Proceedings*, pages 57–59. CEUR-WS.org, 2018.