Computable News Ecosystems: Roles for Humans and Machines

David Caswell

Reynolds Journalism Institute, Missouri School of Journalism, Columbia, MO

Abstract

Two contrasting paradigms for structuring news events and storylines are identified and described: the automated paradigm and the manual paradigm. A specific manual news structuring system is described, and the high-level results of three reporting experiments conducted using the system are presented. In light of these results I then compare automated and manual approaches and argue that they are complementary. A proposal for integrating automated and manual techniques within a structured news ecosystem is presented, and recommendations for integrated approaches are provided.

1 Introduction

News is produced and consumed within local, national and global ecosystems. These ecosystems are made up of large numbers of diverse organizations and individuals, playing a variety of roles: newspapers, news websites, television channels, wire services, aggregators, specialty publishers, sources, freelance reporters, bloggers, social media contributors, advertisers, business and government intelligence organizations, individual news consumers and many others. These ecosystem participants collectively create, remix, exchange, distribute and consume vast quantities of information, almost entirely as discrete blocks of natural language in the form of text articles or scripted video segments.

News ecosystems are currently being disrupted by the internet-driven democratization of publishing and the resulting commodification of text and video (Anderson et al., 2014). This commodification has damaged the economic foundations of news producers, but it has also dramatically increased the depth and breadth of news available to consumers. Unfortunately it is difficult for news consumers to take full advantage of this increased quantity of news, because of the difficulty of converting the information content of large numbers of text articles or video segments into coherent narratives necessary for human understanding (Holton and Chyi, 2012). Overwhelming quantities of text articles and the resulting sense-making challenge for news consumers have been partially addressed by search, personalized article ranking, collaborative filtering, social media curation and other approaches, however, while useful, these technologies have not solved the text overload problem.

A new approach to this sense-making challenge is emerging. It proposes replacing the text article as the primary 'unit of news' with new information artifacts that are aligned with human models of narrative coherence but which are also directly computable. These information artifacts are typically forms of structured 'storylines', in which news events and narratives are recorded as structured representations based on ontologies of semantic frames or event abstractions, each containing well-defined semantic roles. Actual news events can then be instantiated by 'filling in' these semantic roles with references to nodes within knowledge graphs. These structured records of events are then organized into storyline structures that exhibit characteristics of coherent narratives, such as semantic zoom, differential value of events and networked interconnection.

Proceedings of 2nd Workshop on Computing News Storylines, pages 1–8,
Austin, TX, November 5, 2016. ©2016 Association for Computational Linguistics

By structuring news and making it directly available for computation it becomes possible to develop novel news products that are more efficient vehicles for human understanding than corpora of text articles, including summarizations, interactive interfaces, personalized news delivery, query tools, question answering, analytics, etc.

The most common paradigm proposed for this new approach is one of automation, specifically of automated reading (Strassel et al., 2010). This paradigm holds that the source of news events and narratives is in vast corpora and continuous streams of text articles, and that the creation of structured news storylines requires systematically examining those text articles with various natural language processing tools in order to automatically identify and de-duplicate news events across documents, and organize those events into structured storylines based on common locations, characters and entities. The automation paradigm is well-aligned with the current preference within software development for statistical and machine learning approaches to knowledge engineering tasks. This paradigm is also aligned with many previous research projects on computational narratives (Mani, 2013)(Chambers and Jurafsky, 2008)(Zarri, 2009), which often assumed text as the source of narratives and sought the representation of those text narratives as their objective.

This position paper proposes a different paradigm. I describe the operation and evaluation of a manual structured news platform, called 'Structured Stories', and I draw on that experience to examine the potential for human editorial workflows in directly structuring news. I argue that human editorial judgement is essential for creating and maintaining high-quality structured storylines and can be feasibly applied at ecosystem scales. The paper also reviews both paradigms from an ecosystem perspective, and identifies complementary opportunities for both approaches within a computable news ecosystem.

2 Automated Structuring of News

The automated construction of structured storylines from corpora of text articles is challenging because of the absence of any theory of natural language that might formalize the recovery of meaning across documents, or even across sentence boundaries within a single document (Hauser et al., 2002). Without such a theory the event and narrative information contained within natural language text must be extracted indirectly, using an array of Natural Language Processing (NLP) tools. This extraction typically requires identifying discrete news events from references to action within text, capturing those events and their participants as semantic frames and associated semantic roles, de-duplicating the events from other references found in other texts, and then organizing the structured events in context within structured storylines using time, location, common entities or cause-and-effect relationships. None of these steps is trivial and errors compound across all steps, but NLP tools such as frame parsers and named entity recognition (NER) have substantially improved the state of the art.

Automated news reading systems have been attempted in university environments since the 1970s (Cullingford, 1978), however practical, scalable products capable of constructing structured storylines from corpora or streams of news articles have only recently been achieved. Two of these systems, GDELT and EventRegistry (Kwak and An, 2016), are essentially databases of news events structured using relatively coarse ontologies of event abstractions and semantic roles, and supporting only simplified storylines based on time, location or common entities. A third system, called NewsReader, supports more complex storylines, including causal chains, and also includes information about pre- and post-event 'states' within its ontology (Rospocher et al., 2016).

These pioneering systems have successfully demonstrated that it is possible to automatically read large corpora or streams of text news articles, individuate discrete news events, and organize them into explorable structured storylines. These structured storylines are clearly useful and can be deployed at scale, but they convey only a tiny fraction of the information available from storylines conveyed using natural language, and they are also subject to a range of errors. The resolution and quality of these structured events and storylines will improve as the NLP technologies upon which they depend improves, however there may be limits to resolution and quality achievable from statistical NLP techniques (Hovy, 2016), and those limits may be substantially below levels necessary for these structured storylines to replace natural language text as units of news within news ecosystems.

3 Manual Structuring of News

'Structured writing' is an alternative to the automated reading paradigm for generating structured news storylines. The structured writing paradigm holds that the source of news events and narratives is in human editorial judgement - existing understanding that resides in the minds of skilled and informed journalists or analysts. Like automated reading systems, structured writing systems provide an ontology of event abstractions with which to structure news events and assemble storylines, but they require human operators, using dedicated interfaces and tools, to decide which news events to encode as structure, which entities fill the semantic roles within those events, and how those events are organized into storylines. Although they have similar utility, and are based on similar structured representations, structured writing systems differ from automated reading systems in that they exhibit advantages and disadvantages associated with human-centered workflows.

The Structured Stories platform (Caswell, 2015) is an example of a structured writing system. This platform was designed primarily as a knowledge representation system for general news, at a level of semantic granularity substantially finer than representation schemes designed for automated systems, such as NewsReader's Events and Situations Ontology, or GDELT's ontology. The semantic foundation used in the Structured Stories ontology is FrameNet (Baker, 2008), and the additional semantic resolution is added by enabling controlled extension of the FrameNet ontology to form journalistic 'event frames'. Actual news events are then instantiated by assigning knowledge graph references to the semantic roles within these event frames. The platform provides an organization scheme for arranging structured events into structured narratives, including a recursive 'sub-narrative' mechanism to provide semantic zoom and a differential value mechanism for detail management. The resulting structured narratives are assembled from references to events, forming a multi-dimensional graph from common events, common characters, common entities, common locations, etc. Event entry by human reporters is achieved using a simplified sequential user interface that is initiated by the selection of a verb and completed by sequential menu selections that enable the reporter to assign references to the semantic roles, provide time/duration, location, etc. The consumption of structured storylines from the Structured Stories database is enabled by a range of interactive techniques, including timelines, flowcharts, image slideshows, bullet points and text articles generated using natural language generation technology. These interactive techniques are delivered via different user interfaces, including a database management interface, an image-centered interface and a mobile interface.

The feasibility of using human reporters to report directly into the Structured Stories platform was assessed during 2015 and early 2016 in three major reporting projects employing a total of 10 reporters. All of these projects reported real-world news, and reporters were not substantially restricted in what they could choose to report. One was conducted as a stand-alone project with full-time reporters under the guidance of a senior editor (Caswell et al., 2015), one was conducted at a major school of journalism, and one was conducted in the newsroom of a major media company. In aggregate this assessment generated about 120 individual structured storylines, containing about 2300 structured events encoded using about 530 event frames and involving about 1100 different participants (characters, entities, locations, etc.). The level of semantic granularity of the structured events was loosely equivalent to that of the primary events reported in informational articles in a regional newspaper (typically 2-3 events per article), and therefore the assessment produced the event information equivalent of approximately 920 de-duplicated text articles. By comparison the Wall Street Journal produces about 240 articles per day (Meyer, 2016), many of which substantially duplicate events across articles.

This experience of manually reporting structured events into the Structured Stories platform produced several high-level results. The granularity of the event representation scheme was sufficient to capture almost all news events that the reporters wanted to report, suggesting that FrameNet is now relatively comprehensive and that the extension to event frames is relatively practical. All reporters were able to individuate discrete news events, to structure those events appropriately and to assemble storylines from those structured events, however there was wide variability in the ability of reporters to do so. Of the ten reporters three adapted very quickly to the process and became very productive within days, four adapted more slowly, requiring experience and feedback to gradually achieve a moderate level of productivity, and three were unable to adapt and remained at a relatively low level of productivity. Based on post-project interviews it appears that the determining factor in a reporter's ability to adapt to the structuring of news may be their general comfort with abstraction. The actual structuring of events by reporters using the user interface appeared to be relatively easy, typically requiring only 1-2 minutes per event, however the reporting and decision making about events and their semantic roles and characteristics was much more time-consuming. Some reporters described experiencing significant boredom in approaching journalism in this way, and some described a loss of satisfaction in being 'arrangers' of news rather than originators of news using traditional journalistic practices. Nonetheless most reporters saw promise in the technique and thought that it might appeal to some journalists, especially with improvements to the user interface and to the editorial workflow.

Other examples of manual systems for structuring news exist. A major research effort has been ongoing at the BBC since 2010, centered on their News Storyline Ontology (Rissen et al., 2013). This ontology is simpler than the Structured Stories ontology and is intended to be eventually used by reporters as part of regular journalism operations at the BBC, and as part of broader recording of the global news activity by the BBC Monitoring team. Another example is Circa (Coddington, 2015), a San Francisco-based news start-up that was founded in 2010 and closed in 2015 and which used a 10-person editorial team to manually structure journalism into discrete 'atomic units' of news, including events, and assemble those 'atoms' into structured storylines. Other, similar, attempts at manually structuring events exist outside of journalism, including the Nano-publication movement (Mons and Velterop, 2009), which seeks to complement or replace scientific publishing using text papers with much more granular 'nano publications' expressed as RDF triples, from which large-scale networked knowledge structures can be assembled. Nano-publication assumes 'crowdsourced' structuring of research results, in which researchers manually structure their own results. Each of these projects is exploring the feasibility of manually creating and curating repositories of semantically-structured storyline-like information artifacts that originate directly as structure rather than as features extracted from text.

The efficiency of data entry into the Structured Stories platform and into other manual news-structuring systems could be substantially improved in several ways. There is a large and growing body of news that is already available as structured data and which could be mapped into structured events and storylines, including sports news, financial news and increasingly large portions of political news. There are novel techniques based on Controlled Natural Language (Schwitter, 2010) that may enable structured events and storylines to be entered using forms of written language that are more familiar to journalists and analysts. There are also clearly opportunities for at least partial automation of event identification and individuation, using tools and techniques developed for fully-automated news structuring.

4 Comparison

The automatic and manual paradigms for structuring news are complementary. The underlying knowledge representation schemes with which they record structured news events and storylines are broadly similar, and the approaches can therefore be considered as different input mechanisms to a single structured news database. The advantages of each method generally address the disadvantages of the other, suggesting that integrating manual and automated approaches is desirable. Similar approaches, which combine machine learning with human decision-making and oversight, are sometimes called 'human-in-the-loop' (HITL) systems and are increasingly being applied in commercial environments (Bridgwater, 2016).

The primary advantage of manual news structuring is that it enables the application of human editorial judgement in event entry and in the creation and editing of storylines. This has numerous benefits, including the ability to substantially increase the semantic granularity of the represented news events, the ability to avoid and correct errors, the ability to anticipate the needs of consumers, the ability to easily handle unusual events or storyline situations, the ability to handle ambiguity, easily de-duplicate events, etc. The primary disadvantage of manual news structuring is the limited scale at which structuring can be done - i.e. the number and breadth of events and storylines that can be structured - and the lack of consistency with which structuring can be done. Other disadvantages arise from the cultural and workflow challenges inherent in using relatively high-skill human reporters or analysts to perform relatively unsatisfying tasks, as observed in the Structured Stories reporting experiments.

The primary advantage of automated news structuring is the scale of news structuring that can be achieved and the consistency by which that structuring can be done. It is possible, for example, to continually scan the entire global news stream, about 5 million text news articles per day (Wedenberg and Sjberg, 2014), and to detect, de-duplicate and structure major news events for insertion into storylines. The challenges of building and deploying automated news structuring tools are significant, however. There are a series of technical challenges in key tasks, including event detection, assignment of semantic roles, de-duplication of events and organization of structured events into storylines. Each of these can be done, but only with relatively high error rates and at relatively simplistic semantic granularity of the structured events. Storylines produced by automated systems are relatively simplistic and may not be engaging enough for broad communication of news to consumers beyond decision-makers with strong information needs. Furthermore, as the economic basis of professional news organizations erodes, automated news structuring approaches are facing a rapid deterioration in the quality of the corpus of text news articles from which they source news events. This reduction in corpus quality is occurring simultaneously with an increase in the quantity of digital text artifacts, thereby forcing automated systems to detect less semantic signal embedded in more semantic noise.

A comparison of automated and manual approaches to structuring news also reveals differing assumptions about the nature of storylines, and of stories generally. The automated approach loosely assumes that storylines are objective features that already exist in reality (or at least in the source corpus) and that must be found or discovered. The manual approach loosely assumes that storylines are necessarily human-created artifacts, with a human purpose, and that therefore there cannot be a story without an author or authors. This is an important distinction, because it determines whether computational storylines are mechanisms that are primarily useful for search in text corpora or mechanisms that are primarily useful for the storage and communication of human understanding. An automated approach to structuring news is essentially a kind of search engine, albeit one that can deliver unusual and valuable results and therefore aid humans in building understanding. In contrast a news structuring method that is subject to human editing, judgement and oversight could directly accumulate the understanding of skilled and informed journalists and analysts and could refine that understanding over time.

5 Computable News as an ecosystem

It is useful to consider the end-to-end creation, management and use of structured news storylines as an ecosystem, or at least as a highly modular system, because such a view encourages the application of automated or manual techniques as appropriate to the common objective of managing news as structured semantic data rather than as collections of text articles. The possible ecosystem described below (and shown in Figure 1) is hypothetical only regarding the integration of its various components, each of which have already been developed, deployed and evaluated in various stand-alone experimental and commercial systems.

A structured news ecosystem would be centered on a single schema for representing news events and storylines as structure, deployed either as multiple interconnected news databases or possibly as a single centralized news database. A wide variety of sources and methods for capturing news events as structure and for entering them into structured storylines within those databases would be deployed, including manual, semi-automated and entirely automated methods. Manual methods could include sequential user interfaces, such as that used in Structured Stories, controlled natural language interfaces, managed crowdsourcing similar to Wikipedia's editorial process and the use of task marketplaces such as Amazon's Mechanical Turk. Semi-automated methods could include workflows that automatically parse events and semantic roles from text, and presented them to human reporter/analysts for verification. Fully automated methods would include the automated parsing of simple, easily-identified events from text in web corpora, the detection of events in raw structured data and the mapping of events from existing structured event data. These various input methods would be only loosely coupled to the data repository, and new sources of structured news events would be integrated as they were developed.

Regardless of the sources of structured news data, it would be necessary for the resulting structured news database(s) to be under human editorial supervision. This is essential for detecting and responding to errors, for handling unusual events or situations, for enabling the use of events of finer semantic granularity than could be handled automatically, for applying judgement about sub-narratives and detail management, for assessing and managing the various event input methods, and for ensuring the coherence of storylines. The burden of such supervision could be managed using various automated processes and analytical tools while retaining human editorial authority over the overall structured dataset.

The methods of using and consuming structured news events and storylines from a news database would also be varied and modular. Simple interactive interfaces, such as lists, timelines, flowcharts, slideshows, cards, etc. would be necessary, as would basic query and search tools. The available semantic structure would also enable advanced interactive interfaces, especially 'chat bot' interfaces that deliver detailed question-answering. Other advanced interfaces could also be integrated, such as on-demand text articles produced using Natural Language Generations (NLG) tools (Graefe, 2016), automatically generated comics, or on-demand video segments produced using automated video production tools (Kay, 2015).

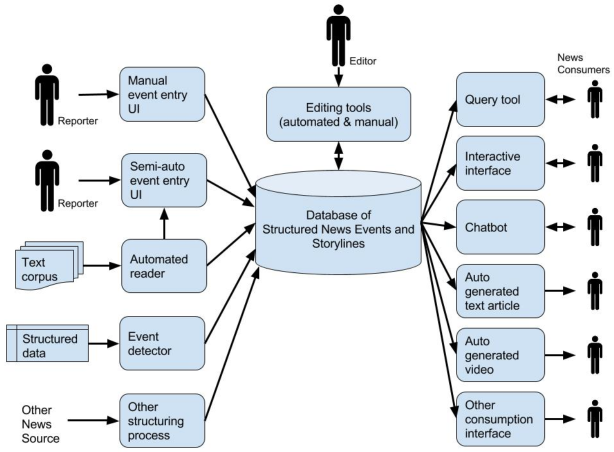

Figure 1: An integrated structured news ecosystem.

All of these components already exist and have been deployed in experimental and commercial systems, therefore the technical challenges of assembling an integrated manual/automated structured news ecosystem would primarily involve integration. The human factor and cultural challenges would probably be more difficult, however there is substantial motivation within existing news organizations to find new ways of automating, bundling and exploiting news. These organizations already have a trained workforce of many tens of thousands of skilled reporters and editors, and it appears from the Structured Stories reporting experiments that portions of this workforce may have the analytical and abstraction skills necessary to transition to new 'meta-journalism' or 'meta-editorial' roles within a structured news environment.

6 Conclusions

The existing text-based news ecosystem is failing for both producers and consumers of news, and novel structured and automated approaches to news are required. The assumption that news events and storylines originate in natural language text and that automated reading is the sole method available to access and structure those events and storylines is limiting. Lessons from the Structured Stories reporting experiments, and from other manual news structuring projects, have shown that applying human editorial judgement to structured news environments is feasible, and can potentially address some of the weaknesses of fully automated systems. Integration of automated and manual approaches to structuring news could enable a structured news ecosystem that exhibits the advantages of each method.

Facilitating the development of a structured news ecosystem requires viewing computable news functionality as modular, with manual, semi-automated and automated components. Integration of these modular components within an ecosystem will require standards - particularly a standard representation schema for structured news events and storylines, and standard interfaces for event entry modules and for storyline consumption modules. Addressing the human factors challenges of an integrated structured news ecosystem will require development of the abstraction skills of reporters, editors and analysts and enabling journalists to practice their profession at a higher level of abstraction. These are not simple challenges, however the experience of building and evaluating both automated and manual news structuring systems has demonstrated that they are achievable.

Acknowledgments

The author would like to acknowledge support for the Structured Stories project from the Reynolds Journalism Institute at the Missouri School of Journalism, The Reporter's Lab at Duke University and the Online News Association Challenge Fund.

References

C.W. Anderson, Emily Bell, and Clay Shirky. 2014. Post industrial journalism: Adapting to the present. Technical report, Tow Center for Digital Journalism, New York, NY.

Collin Baker. 2008. Framenet, present and future. *Proceedings of the First International Conference on Global Interoperability for Language Resources*.

Adrian Bridgwater. 2016. Machine learning needs a human-in-the-loop. *Forbes*, March.

David Caswell, Frank Russell, and Bill Adair. 2015. Editorial aspects of reporting into structured narratives. *Proceedings of the 2015 Computation + Journalism Symposium*.

David Caswell. 2015. Structured narratives as a framework for journalism: A work in progress. *Proceedings of the Sixth International Workshop on Computational Models of Narrative*.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797. Association for Computational Linguistics.

Mark Coddington. 2015. *Telling Secondhand Stories: News Aggregation and the Production of Journalistic Knowledge*. Ph.D. thesis, The University of Texas at Austin, Austin, TX.

Richard Edward Cullingford. 1978. Script application: Computer understanding of newspaper stories. Technical report, Yale University, New Haven, CT.

Andreas Graefe. 2016. Guide to automated journalism. Technical report, Tow Center for Digital Journalism, New York, NY.

Marc Hauser, Noam Chomsky, and Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298:1569–1579.

Avery E. Holton and Hsiang Iris Chyi. 2012. News and the overloaded consumer: Factors influencing information overload among news consumers. *Cyberpsychology, Behavior, and Social Networking*, 15:619–624, November.

Eduard Hovy. 2016. Filling the long tail. Keynote presentation of the 2nd Spinoza Workshop: Looking at the Long Tail, Vrije Universiteit Amsterdam, June.

Hilary Kay. 2015. This is how text-to-video technology works. Technical report, Wibbitz Inc.

Haewoon Kwak and Jisun An. 2016. Comparison of widely used world news datasets: Gdelt and eventregistry. *Proceedings of the 23rd International Conference on Web and Social Media*.

Inderjeet Mani. 2013. *Computational Modeling of Narrative*. Morgan and Claypool, San Rafael, CA.

Robinson Meyer. 2016. How many stories do newspapers publish per day? *The Atlantic*.

Barend Mons and Jan Velterop. 2009. Nano-publication in the e-science era. *Workshop on Semantic Web Applications in Scientific Discourse*.

Paul Rissen, Helen Lippell, Matt Chadburn, Tom Leitch, Dan Brickley, Michael Smethurst, and Sebastien Cevey. 2013. News storyline ontology.

Marco Rospocher, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. 2016. Building event-centric knowledge graphs from news. *Web Semantics: Science, Services and Agents on the World Wide Web*, 37:Pages 132–151.

Rolf Schwitter. 2010. Controlled natural languages for knowledge representation. *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1113–1121.

Stephanie Strassel, Dan Adams, Henry Goldberg, Jonathan Herr, Ron Keesing, Daniel Oblinger, Heather Simpson, Robert Schrag, and Jonathan Wright. 2010. The darpa machine reading program - encouraging linguistic and reasoning research with a series of reading tasks. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), may.

Kim Wedenberg and Alexander Sjberg. 2014. Online inference of topics: Implementation of lda topic modeling using an online variational bayes inference algorithm to sort news articles. Technical report, Uppsala Universitet, Uppsala, Sweden, February.

Gian Piero Zarri. 2009. *Representation and Management of Narrative Information*. Springer-Verlag, London, United Kingdom.