Finding Stories in 1,784,532 Events: Scaling Up Computational Models of Narrative*

Marieke van Erp, Antske Fokkens, and Piek Vossen

The Network Institute
VU University Amsterdam
Amsterdam, the Netherlands
{marieke.van.erp,antske.fokkens,piek.vossen}@vu.nl

Abstract

Information professionals face the challenge of making sense of an ever increasing amount of information. Storylines can provide a useful way to present relevant information because they reveal explanatory relations between events. In this position paper, we present and discuss the four main challenges that make it difficult to get to these stories and our first ideas on how to start resolving them.

1998 ACM Subject Classification H.3.3. Information Search and Retrieval

Keywords and phrases big data, news, aggregation, story detection

Digital Object Identifier 10.4230/OASIcs.CMN.2014.241

1 Introduction and Motivation

Every working day, millions of news articles are produced by thousands of different sources that report on many different events that happened, are happening, or may or will happen in the world. Some sources provide the same account, some complement each other, some provide different perspectives and some sources contradict each other. Information professionals are facing the challenge of making sense of this ever increasing information deluge [8].

A core task here is to create reconstructions of what happened where, when and to whom, to provide explanations of particular events or identify relevant actors. They can be seen as narratives or stories about the real world that need to be discovered in the data. We consider storylines the most compact and informative structures for representing the essence of large volumes of news data over longer periods of time summarising the changes reported in the news as sequences of events involving participants but also hinting at explanations and point to the forces at work.

In the NewsReader project,[1] we aim to support information professionals by automating the reconstruction of storylines from large amounts of news articles over longer periods of time. We are developing natural language processing technology to process daily news streams in four languages (English, Spanish, Italian and Dutch) to extract events and their arguments. Whilst we have a clear idea of how to represent this information as structured events, we are still investigating how to go beyond these events to automatically detect and represent storylines from literally tons of sources.

For most work on narrative analysis and modelling, the unit of the story is known e.g. a folk tale [11], novel [7], or film [4]. Identifying stories in large amounts of newspaper texts

* This research was funded by the European Union's 7th Framework Programme via the NewsReader (ICT-316404) project.
[1] http://www.newsreader-project.eu

5th Workshop on Computational Models of Narrative (CMN'14).
Editors: Mark A. Finlayson, Jan Christoph Meister, and Emile G. Bruneau; pp. 241–245

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

introduces several novel challenges. In this position paper, we outline the challenges that are involved when storylines are hidden in a large set of events coming from different sources.

This paper is organised as follows. In Section 2, we describe the global automotive industry use cases from the NewsReader project. In Section 3, we describe the four main challenges we have identified: scope, granularity, identity and change, and perspective. We summarise our main findings in Section 4.

2 Global automotive industry

The global automotive industry provides us with a case that is rich in complex and varied events, interactions between key players and possible storylines. For a first pilot, we have collected and processed 64,423 news articles from the LexisNexis archive[2] (out of their 6.1 million English available news articles about the car industry between 2003 and 2013). The processing consisted of a traditional natural language processing pipeline in which we first performed a structural analysis of each article (e.g. part-of-speech tagging, syntactic parsing) of the text followed by a semantic analysis (e.g. named entity recognition and linking, event detection and semantic-role-labeling, opinion mining) [1]. Then, mentions of events and entities are matched across the documents to establish coreference relations following [5], i.e. what accounts write about the same events. The coreference relations are used to aggregate information for uniquely defined instances of entities and events. This results in a reduction from 3,127,446 textual mentions of actors (for example persons or companies) to 445,286 instances. For locations, there is a reduction from 1,049,711 textual mentions to 62,255 instances and for events a reduction from 5,247,872 text mentions to 1,784,532 instances.

We know a host of interesting stories are to be found inside this collection of events, actors and locations. There is for example the story of the Porsche and Volkswagen take-over that describes Porsche buying an ever growing stake in Volkswagen between 2005 and 2009, prompting speculations that Porsche would take over Volkswagen. However, with the turn of the economy, Volkswagen reversed the tide in 2009 and eventually took over Porsche. Around this story, there are related stories revolving around the different actors related to these companies such as those of key staff members Wendelin Wiedeking and Ferdinand Piëch. These stories are only the tip of the iceberg. Through automatic detection of interesting stories we aim to discover new stories that were previously hidden in the data. The main challenges we encounter in identifying these stories are presented in Section 3.

3 Narrative Detection and Modelling at Scale

While each news article tells a story on its own, we aim to construct a story across different news articles that may be published over long timespans. We have identified four challenges that come into play when information comes from different sources which may not all be related: scope, granularity, identity and change, and perspectives.

3.1 Scope

When considering news articles around a certain time, we do not know a priori which stories are interesting or relevant to tell, who the relevant actors (characters) are in our domain and which events are relevant to include in the story. We thus need to determine the scope or

[2] http://www.lexisnexis.com

extent of the story. As we have a continuous stream of incoming news articles, one of the things we also need to identify where a story begins and where it ends.

Most news aggregation systems[3] and story detection approaches from news streams such as [10] use some measure of frequency within a particular timeframe to assess the trendiness of a topic which may indicate its importance, as well as its rise and fall. While this may be a good measure for identifying the main events in popular stories, it is not suitable for identifying less popular stories. These techniques may thus not lead to novel insights but rather point to (parts of) stories that are most well-known.

Besides events and characters being mentioned frequently (the volume of the news), we are therefore also looking into detecting strong emotions and opinions (sentiment analysis), impact or involvement of a larger number of characters (local vs global reporting) and particular events types or actors that are valued (for example particular scenarios or, following [2], interesting or key characters in the domain).

Once the events around the climax of a story are identified, events that led up to the climax can be traced. Temporal and causal relations between events, involved actors, intentions and speculations (e.g. speculations on a take-over, a promise by a CEO) and critical changes of states (e.g. a series of take-overs eventually resulting in a company becoming a market leader) can be used to trace these events. We use our NLP analysis tools to detect relations, entities and events, together with domain knowledge to type events, identify motifs or scenarios and importance of events in a particular scenario. We can utilise coherence measures as proposed in [12] to ensure the selected events form a coherent story.

3.2 Granularity

As the goal of NewsReader is to support information specialists in reconstructing stories relevant to their domain, we interviewed some information specialists. We learnt from these interviews that their daily work includes high-level abstract cases as well as fine-grained detailed cases and everything in between. Our story model therefore needs to be able to detect and model both high and low level stories. Users may initially start with a higher level story, and then find that they need to zoom in on particular details. It is therefore important to be able to identify high-level stories such as the Volkswagen-Porsche take-over, but also its finer grained events (e.g. changes in the stock market), actors (e.g. persons involved in negotiations) and storylines (e.g. developments within Porsche) are part of a more general storyline.

Modelling hierarchical relations between entities that enable us to switch between for example corporate-level and person-level stories may be a first step towards addressing this challenge. Such hierarchies may also come in useful for the next challenge.

3.3 Identity and Change

News stories report on changes in the world. While events are identified as points of change, these often change the state of the actors involved too. In the car domain, actors merge or take over each other and some actors are thus absorbed. This happened for example to Daewoo, a South Korean car maker that has gone through several name changes since its founding in 1982. In 2001, General Motors took over Daewoo and in 2005, the names of all Daewoo car models were re-badged as Chrevrolet models in Europe, thus effectively making

[3] http://emm.newsexplorer.eu/NewsExplorer/home/en/latest.html

the name 'Daewoo' disappear on this continent. If one were to do a simple search on the database for events in which Daewoo was involved, the events would stop after 2005, but one could also argue that they continued under a different label.

Detecting and modelling the relationships between different actors through time is necessary to deal with this from a database querying perspective. From a storyline perspective, one needs to be able to express hierarchies of actors (for subsidiaries of different companies for example) as well as actor changes through time. We are investigating whether some advances made in the knowledge representation domain can be applied to NewsReader storylines. As a starting point we take [9] which presents a model for capturing entity changes over time.

3.4 Perspective

It is commonly accepted in the theory of narratology that there is a distinction between the story itself (the fabula) and way the story is told (discourse) [3]. News sources are not objective [13] as there are various reasons for a source to present an event in the way it does, or even to select an event for presentation.

As the NewsReader project analyses news articles from a large variety of sources (3,111 for the cars data), it is inevitable that some of these sources contradict each other. In a recent car recall case, one source states that 900,000 cars are recalled,[4] while another claims 'only' 644,000 cars are recalled.[5] We currently use a similarity measure between event descriptions to establish whether two sources are talking about the same event. This makes it particularly challenging to automatically detect when contradictory information points to different events and when contradictory information provides alternative perspectives.

However, we have developed a model to represent such information. Through the Grounded Annotation Framework [6],[6] we can store the source of each event mention. This allows us to compare how different sources talk about a specific event. We can look into the number of references to an event or amount of detail a specific source provides. However, in order to detect more subtle differences in perspectives, we will need to do a deep analysis of the text and look for example at stylistic differences such as word use and differences in focus of the article or level of detail given.

4 Conclusion

We presented the challenges encountered when one tries to identify stories in large amounts of data consisting of many units that may or may not be related to the same story. Our four main challenges related to this goal are: scope, granularity, identity and change, and perspectives. We presented our current research directions for identifying the scope of a story and referred to GAF which allows us to compare perspectives. Dealing with granularity, identity change and identifying alternative perspectives are still open challenges.

[4] http://www.dallasnews.com/business/business-headlines/20140402-chrysler-recalls-nearly-900000-jeeps-durangos-to-fix-brake-problem.ece Retrieved 4 April 2014

[5] http://www.autoblog.com/2014/04/02/chrysler-recall-644k-jeep-grand-cherokee-dodge-durango-brakes-official/ Retrieved 4 April 2014

[6] http://groundedannotationframework.org

References

1 Rodrigo Agerri, Josu Bermudez, and German Rigau. Ixa pipeline: Efficient and ready to use multilingual nlp tools. In *To appear in Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31 2014.

2 Julio César Bahamón and R. Michael Young. A choice-based model of character personality in narrative. In *Proceedings of the 2012 Workshop on Computational Models of Narrative*, Istanbul, Turkey, May 26-27 2012.

3 Mieke Bal. De theorie van verhalen en vertellen. *Inleiding in de narratologie*, 1980.

4 David Bordwell. *Narration in the Fiction Film*. Routledge, 1985.

5 Agata Cybulska and Piek Vossen. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *To appear in Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31 2014.

6 Antske Fokkens, Marieke van Erp, Piek Vossen, Sara Tonelli, Willem Robert van Hage, Luciano Serafini, Rachele Sprugnoli, and Jesper Hoeksema. GAF: A grounded annotation framework for events. In Eduard Hovy, Teruko Mitamura, and Martha Palmer, editors, *Proceedings of the 1st workshop on Events: Definition, Detection, Coreference, and Representation at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL2013)*, Atlanta, GA, USA, Jun 9-15 2013. Association for Computational Linguistics.

7 Gérard Genette. *Narrative Discourse*. Basil Blackwell, 1980.

8 Jonathan Gray, Liliana Bounegru, and Lucy Chambers, editors. *The Data Journalism Handbook*. O'Reilly Media, 2012.

9 Harry Halpin and Patrick J. Hayes. When owl:sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web. In *Proceedings of LDOW2010*, Raleigh, NC, USA, April 27 2010.

10 Saša Petrović, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to twitter. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL2010)*, 2010.

11 Vlad'imir Propp. *Morphology of the Folk Tale*. The American Folklore Society and Indiana University, 2nd edition, 1968.

12 Dafna Shahaf and Carlos Guestrin. Connecting the dots between news articles. In *Proceedings of KDD'10*, Washington, DC, USA, July 25-28 2010.

13 Andrea Umbricht. Patterns of news making in western journalism a content analysis of newspapers across six western democracies and five decades. Technical report, Institute of Mass Communication and Media Research (IPMZ), 2014.