Different Types of Automated and Semi-automated Semantic Storytelling: Curation Technologies for Different Sectors

Georg Rehm[1(✉)], Julián Moreno-Schneider[1], Peter Bourgonje[1], Ankit Srivastava[1], Rolf Fricke[2], Jan Thomsen[2], Jing He[3], Joachim Quantz[3], Armin Berger[4], Luca König[4], Sören Räuchle[4], Jens Gerth[4], and David Wabnitz[5]

[1] Language Technology Lab, DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany
georg.rehm@dfki.de
[2] Condat GmbH, Alt-Moabit 91c, 10559 Berlin, Germany
[3] 3pc GmbH Neue Kommunikation, Prinzessinnenstraße 1, 10969 Berlin, Germany
[4] ART+COM AG, Kleiststraße 23-26, 10787 Berlin, Germany
[5] Kreuzwerker GmbH, Ritterstraße 12-14, 10969 Berlin, Germany
http://digitale-kuratierung.de

Abstract. Many industries face an increasing need for smart systems that support the processing and generation of digital content. This is both due to an ever increasing amount of incoming content that needs to be processed faster and more efficiently, but also due to an ever increasing pressure of publishing new content in cycles that are getting shorter and shorter. In a research and technology transfer project we develop a platform that provides content curation services that can be integrated into Content Management Systems, among others. In the project we develop curation services, which comprise semantic text and document analytics processes as well as knowledge technologies that can be applied to document collections. The key objective is to support digital curators in their daily work, i.e., to (semi-)automate processes that the human experts are normally required to carry out intellectually and, typically, without tool support. The goal is to enable knowledge workers to become more efficient and more effective as well as to produce high-quality content. In this article we focus on the current state of development with regard to semantic storytelling in our four use cases.

1 Introduction

Digital content and online media have reached an unprecedented level of relevance and importance, especially with regard to commercial but also political and societal aspects. One of the many technological challenges refers to better support and smarter technologies for digital content curators, i.e., persons, who work primarily at and with a computer, who are facing an ever increasing incoming stream of heterogeneous information and who create, in a general

G. Rehm and T. Declerck (Eds.): GSCL 2017, LNAI 10713, pp. 232–247, 2018.
https://doi.org/10.1007/978-3-319-73706-5_20

sense, new content based on the requirements, demands, expectations and conventions of the sector they work in. For example, experts in a digital agency build websites or mobile apps for clients who provide documents, data, pictures, videos and other assets that are processed, sorted, augmented, arranged, designed, packaged and then deployed. Knowledge workers in a library digitise a specific archive, add metadata and critical edition information and publish the archive online. Journalists need to stay on top of the news stream including blogs, microblogs, newswires etc. in order to produce a new article on a breaking topic. A multitude of examples exist in multiple sectors and branches of media (television, radio, blogs, journalism etc.). All these different professional environments can benefit immensely from semantic technologies that support knowledge workers, who typically work under high time pressure, in their activities: finding relevant information, highlighting important concepts, sorting incoming documents, translating articles in foreign languages, suggesting interesting topics etc. We call these different semantic services, that can be applied flexibly in different professional environments that all have to do with the processing, analysis, translation, evaluation, contextualisation, verification, synthesis and production of digital information, *Curation Technologies*.

The activities reported in this paper are carried out in the context of a two-year research and technology transfer project, Digital Curation Technologies[1] in which DFKI collaborates with four SME companies that operate in four sectors (3pc: public archives; Kreuzwerker: print journalism; Condat: television and media; ART+COM: museum and exhibition design). We develop, in prototypically implemented use cases, a flexible platform that provides generic curation services such as, e.g., summarisation, named entity recognition, entity linking and machine translation (Bourgonje et al. 2016a,b). These are integrated into the partners' in-house systems and customised to their domains so that the content curators who use these systems can do their jobs more efficiently, more easily and with higher quality. Their tasks involve processing, analysing, skimming, sorting, summarising, evaluating and making sense of large amounts of digital content, out of which a new piece of digital content is created, e.g., an exhibition catalogue, a news article or an investigative report.

We mainly work with self-contained document collections but our tools can also be applied to news, search results, blog posts etc. The key objective is to shorten the time it takes digital curators to familiarise themselves with a large collection by extracting relevant data and presenting the data in a way that enables the user to be more efficient, especially when they are not domain experts.

We develop modular language and knowledge technology components that can be arranged in workflows. Based on their output, a semantic layer is generated on top of a document collection. It contains various types of metadata as annotations that can be made use of in further processing steps, visualisations or user interfaces. Our approach bundles a flexible set of semantic services for the *production of digital content*, e.g., to recommend or to highlight interesting

[1] http://digitale-kuratierung.de.

and unforeseen storylines or relations between entities to human experts. We call this approach *Semantic Storytelling*.

In this article we concentrate on the collaboration between the research partner and the four SME companies. For each use case we present a prototype application, all of which are currently in experimental use in these companies.

2 Curation Technologies

The curation services are made available through a shared platform and RESTful APIs (Bourgonje et al. 2016a; Moreno-Schneider et al. 2017a; Bourgonje et al. 2016b; Srivastava et al. 2016). They comprise modules that either work on their own or that can be arranged as workflows.[2] The various modules analyse documents and extract information to be used in content curation scenarios. Interoperability between the modules is achieved through the NLP Interchange Format (NIF) (Sasaki et al. 2015). NIF allows for the combination of web services in a decentralised way, without hard-wiring specific pipelines. In the following we briefly present selected curation services.

2.1 Named Entity Recognition and Named Entity Linking

First we convert every document to NIF and then perform Named Entity Recognition (NER). NER consists of two different approaches that allow training with annotated data and/or to use dictionaries. Afterwards the service attempts to look up any named entity in its (language-specific) DBpedia page using DBpedia Spotlight (2016) to extract additional information using SPARQL.

2.2 Geographical Localisation Module and Map Visualisations

The geographical location module uses SPARQL and the Geonames ontology (Wick 2015) to retrieve the latitude and longitude of a location as specified in its DBpedia entry. The module also computes the mean and standard deviation value for latitude and longitude of all identified locations in a document. With this information we can position a document on a map visualisation.

2.3 Temporal Expression Analysis and Timelining

The temporal expression analyser consists of two approaches that can process German and English natural language text, i.e. a regular expression grammar and a modified implementation of HeidelTime (Strötgen and Gertz 2013). After identification, temporal expressions are normalised to a shared format and added to the NIF representation to enable reasoning over temporal expressions and also for archiving purposes. The platform adds document-level statistics based on normalised temporal values. These can be used to position a document on a timeline.

[2] Moreno-Schneider et al. (2017a) describes the Semantic Storytelling curation service and provides more technical details. The platform itself is based on the FREME infrastructure (Sasaki et al. 2015).

2.4 Text Classification and Document Clustering

We provide a generic classification service, which is based on Mallet (McCallum 2002). It assigns topics or domains such as "politics" or "sports" to documents when labeled training data is available. Annotated topics are stored in the NIF representation as RDF. Unsupervised document clustering is performed using the Gensim toolkit (Řehůřek and Sojka 2010). For the purpose of this paper we performed experiments with a bag-of-words approach and with tf/idf transformations for the Latent Semantic Indexing (LSI) (Halko et al. 2011), Latent Dirichlet Allocation (LDA) (Hoffman et al. 2010) and Hierarchical Dirichlet Process (HDP) (Wang et al. 2011) algorithms.

2.5 Coreference Resolution

For the correct interpretation and representation of events and their arguments and components, the resolution of mentions referring to entities that are not identified by the NER component (because they are realised by a pronoun or alternative formulation) is essential. For these cases we implemented a coreference resolution mechanism based on CoreNLP for English (Raghunathan et al. 2010). For German language documents we replicated this multi-sieve approach (Srivastava et al. 2017). This component increases the coverage of the NER and event detection modules.

2.6 Monolingual and Cross-Lingual Event Detection

We implemented a state-of-the-art event detection system based on Yang and Mitchell (2016) to pinpoint words or phrases in a sentence that refer to events involving participants and locations, affected by other events and spatio-temporal aspects. The module is trained on the ACE 2005 data (Doddington et al. 2004), consisting of 529 documents from a variety of sources. We apply the tool to extract generic events from the various datasets in our curation scenarios. We also implemented a cross-lingual event detection system, i.e., we translate non-English documents to English through Moses SMT (Koehn et al. 2007) and detect events in the translated documents using the system described above.

2.7 Single and Multi-document Summarisation

Automatic summarisation refers to reducing input text (from one or more documents) into a shorter version by keeping its main content intact while still conveying the actual desired meaning (Ou et al. 2008; Mani and Maybury 1999). This task typically involves identifying, extracting and reordering the most important sentences from a document (collection) into a summary. We offer three different approaches: centroid-based summarisation (Radev et al. 2000), lexical page ranking (Erkan and Radev 2004), and cluster-based link analysis (Wan and Yang 2008).

2.8 User Interaction in the Curation Technologies Prototypes

Our primary goal is to support knowledge workers by automating some of their typical processes. This is why all implemented user interfaces are inherently interactive. By providing feedback to, for example, the output of certain semantic services, knowledge workers have some amount of control over the workflow. They are also able to upload existing resources to adapt individual services. For example, we allow users to identify errors in the output (e.g., incorrectly identified entities) and provide feedback to the algorithm; NER allows users to supply dictionaries for entity linking; Event Detection allows users to supply lists of entities for the identification of agents for events.

3 Semantic Storytelling: Four Sector-Specific Use Cases

Generic Semantic Storytelling involves processing a coherent and self-contained collection of documents in order to identify and to suggest, to the human curator, on a rather abstract level, one or more potential story paths, i.e., specific relationships between entities that can then be used for the process of structuring a new piece of content. It was a conscious decision not to artificially restrict the approach (for example, to certain text types) but to keep it broad and extensible so that we can apply it to the specific needs and requirements of different sectors. In one sector a single surprising, hitherto unknown relation between two entities may be enough to construct an actual story while in others we may try to generate the base skeleton of a storyline semi-automatically (Moreno-Schneider et al. 2016). One concrete example are millions of leaked documents, in which an investigative journalist wants to find the most interesting nuggets of information, i.e., surprising relations between different entities, say, politicians and offshore banks. Our services do not necessarily have to exhibit perfect performance because humans are always in the loop in our application scenario. We want to provide robust technologies with broad coverage. For some services this goal can be fulfilled while for others, it is a bit more ambitious.

3.1 Sector: Museums and Exhibitions

The company ART+COM AG is specialised in the design of museums, exhibitions and showrooms. Their creative staff needs to be able to familiarise themselves with new topics quickly to participate in pitches or during the execution of projects. We implemented a graphical user interface (GUI) that supports the knowledge workers' storytelling capabilities, e.g., for arranging exhibits in a room or for arranging the rooms themselves, by supporting and improving the task of curating incoming content. The GUI enables the effective interaction with the content and the semantic analysis layer. Users can get a quick overview of a specific topic or drill down into the semantic knowledge base to explore deeper relationships.

Initial user research provided valuable insights into the needs of the knowledge workers in this specific use case, especially regarding the kinds of tools

and environments each user is familiar with as well as extrapolating their usage patterns (Rehm et al. 2017a). Incoming content materials, provided by clients, include large heterogeneous document collections, e.g., books, images, scientific papers etc. We subdivide the curation process into the phases search, evaluate, organise.

The prototype is a web application (Fig. 1). Users can import documents, such as briefing materials from the client, or perform explorative web searches. Content is automatically analysed by the curation services. The application performs a lookup on the extracted information, e.g., named entities, on Wikidata in order to enrich the entities with useful additional information. Entities are further enriched with top-level ontology labels in order to provide an overview of the distribution of information in categories, for instance, person, organisation, and location. Intuitive visualisation of extracted information is a focus of this prototype. We realised several approaches including a network overview, semantic clustering, timelining and maps. In an evaluation the knowledge workers concluded that the implemented interfaces provides a good overview of the subject and that they would use the tool at the beginning of a project, particularly when confronted with massive amounts of text (Rehm et al. 2017a).

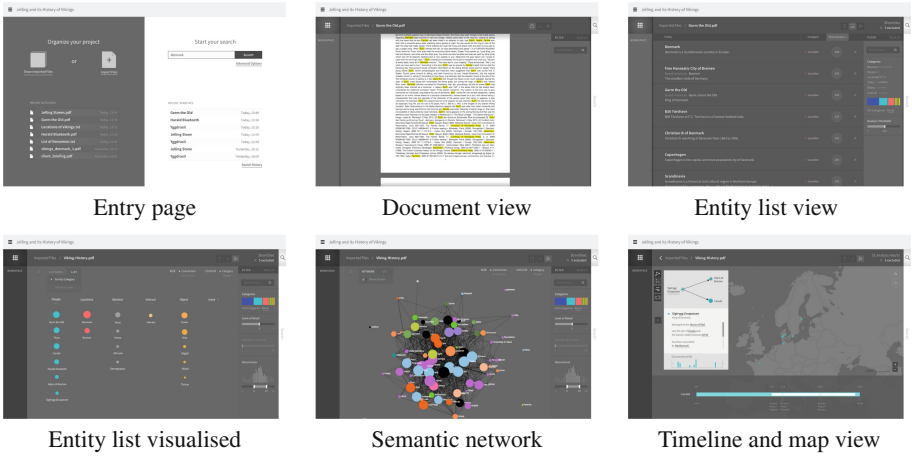

Fig. 1. Prototype application for the sector museums and exhibitions

3.2 Sector: Public Archives, Libraries, Digital Humanities

For the company 3pc GmbH we developed an authoring environment, enabled by the curation technology platform. Many of 3pc's projects involve a client, e.g., a company, an actress or a political party, that provides a set of digital content and a rough idea how to structure and visualise these assets in the form of a website or app. A tool that can semantically process such a document collection to enable the efficient authoring of flexible, professional, convincing,

visually appealing content products that provide engaging stories and that can be played out in different formats (e.g., web app, iOS or Android app, ebook etc.) would significantly reduce the effort on the side of the agency and improve their flexibility. Several screens of the authoring environment's GUI are shown in Fig. 2. It was a conscious design decision to move beyond the typical notion of a "web page" that is broken up into different "modules" using templates. The focus of this prototype are engaging stories told through informative content.

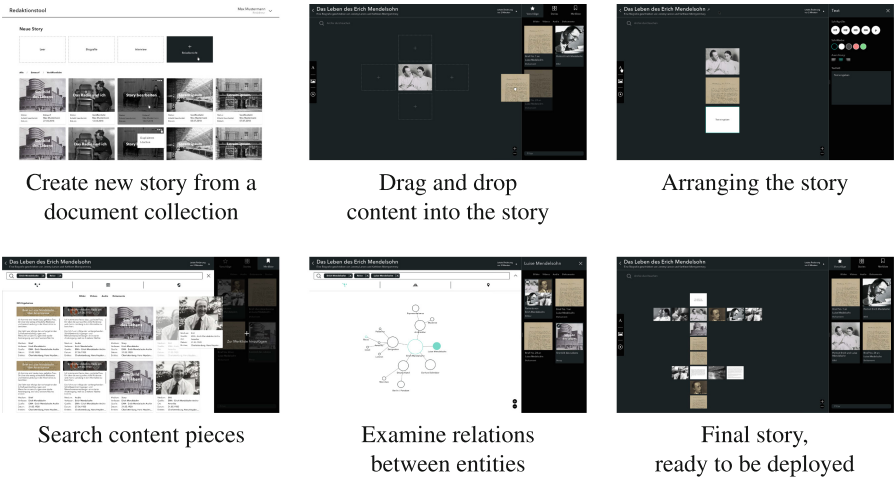

Fig. 2. Prototype application for the sector archives and libraries

With this tool the content curator can interactively put together a story based on the semantically enriched content. In the example use case we work with a set of approx. 2,800 letters exchanged between the German architect Erich Mendelsohn (1887–1953) and his wife Luise, both of whom travelled frequently. The collection contains 2,796 letters, written between 1910 and 1953, with a total of 1,002,742 words (359 words per letter on average) on more than 11,000 sheets of paper. Most are in German (2,481), the rest is written in English (312) and French (3). The letters were scanned, transcribed and critically edited; photos and metadata are available; this research was carried out in a project that the authors of the present paper are not affiliated with (Bienert and de Wit 2014). In the letters the Mendelsohns discuss their private and professional lives, their relationship, meetings with friends and business partners, and also their travels.

We decided to focus upon identifying all movement action events (MAE), i.e., all trips undertaken by a subject (usually the letter's author) from location A to location B with a specific departure and arrival time, using a specific mode of transport. This way we want to transform, ideally automatically, a set of interconnected *letters* into a *travelogue* that provides an engaging story to the reader and that also enables additional modes of access, e.g., through map-based or timeline-based visualisations. We wanted to explore to what extent it is

possible to automate the production of an online version of such a collection. A complete description can be found in (Rehm et al. 2017b); here we only present a few examples of extracted MAEs to demonstrate the functionality (Table 1). An MAE consists of the six-tuple MAE = $<P, L_O, L_D, t_d, t_a, m>$ with P a reference to the participant (E. or L. Mendelsohn), L_O and L_D references to the origin and destination locations (named locations, GPS coordinates), t_d and t_a the time of departure and arrival and m the mode of transport. Each component is optional as long as the MAE contains at least one participant and a destination.

Table 1. Automatically extracted movement action events (MAEs)

Letter text	Extracted MAEs
Another train stopped [...] this would be the train with which Eric had to leave Cleveland	Eric, Cleveland, [], [], [], train
Because I have to leave on the 13th for Chicago	I (Erich), Croton on Hudson, NY, Chicago, 13th Dec. 1945, [], []
April 5th 48 Sweetheart - Here I am - just arrived in Palm Springs [...]	I (Erich), [], Palm Springs, [], 5th April 1948, []
Thompsons are leaving for a week - [...] at the Beverly Hills on Thursday night!!	Thompsons, [], Beverly Hills, 8th July, [], []

3.3 Sector: Journalism

Journalists write news articles based on information collected from different sources (news agencies, media streams, other news articles, sources, etc.). Research is needed on the topic and domain at hand to produce a high-quality piece. Facts have to be checked, different view points considered, information from multiple sources combined in a sensible way. The resulting piece usually combines new, relevant and surprising information regarding the event reported upon. While the amount of available information is increasing on a daily basis, the journalist's ability to go through all the data is decreasing, which is why smart technology support is needed. We want to enable journalists interactively to put together a story based on semantic content enrichment. In our various use cases, different parts of the content function as atomic building blocks (sentences, paragraphs, documents). For this use case we focus, for now, upon document-level building blocks for generating stories, i.e., documents can be rearranged, included and deleted from a storyline.[3]

[3] In a follow-up project we plan to use smaller content components with which we will experiment towards the generation of articles based on multiple story paths, automatically generated with the help of semantic annotations.

For the company Kreuzwerker GmbH we developed an extension for the open source newsroom software Superdesk (https://www.superdesk.org). This production environment specialises on the creation of content, i.e., the actual play-out and rendering of the content is taken care of by other parts of a larger system. The plug-in allows the semantic processing of incoming news streams to enable smart features, e.g., keyword alerts, content exploration, identifying related content, summarisation and machine translation. It also allows for the visualisation and annotation of news documents using additional databases and knowledge graphs (e.g., Linked Open Data) to enable faceted search scenarios so that the journalist has fine-grained mechanisms to locate the needle in a potentially very large digital haystack. Faceted search includes entities, topics, sentiment values and genres, complemented with semantic information from external sources (DBpedia, WikiData etc.). Menus show the annotated entities and their frequencies next to a set of related documents. Example screens of this newsroom content curation dashboard are shown in Fig. 3. The plug-in mainly operates on the (1) ingest view and the (2) authoring view. The first view allows to ingest content channels into the production environment; the semantic tools (see Sect. 2) can automatically analyse the content using, e.g., topic detection, classification (e.g., IPTC topics) and others. In the second view, the curation tools support the authoring process, to add or modify annotations and to recommend related content. A thorough description of the use case and the developed prototype can be found in (Moreno-Schneider et al. 2017b).

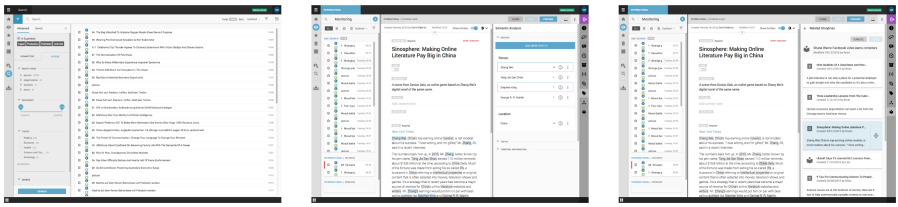

Fig. 3. Prototype application for the sector journalism

The company Condat AG develops software that is used in television stations, supporting journalists to put together, e.g., news programmes, by providing access to databases of metadata-rich archives that consist of media fragments. We developed recommendation, metadata extraction and multi-document summarisation services that enable editors to process large amounts of data, to find updates of stories about events already seen and to identify relevant, but rarely used, media, to provide a certain level of surprise in the storytelling.

Current news exploration systems such as, e.g., Google News, rely on the extraction of entities and analytics processes that operate on frequencies and timestamps to populate categories that resemble traditional newspaper sections (National, World, Politics, Sports, etc.). It also includes “highlight news”, which

consists of named entities. At the time of writing, "London" was part of the "highlight news" but, as a label, it is not helpful – "Brexit" or "Grenfell Tower fire" would have been more appropriate. Based on simple entity frequencies we cannot distinguish between news about these independent events. We attempt to group documents from a newsstream, based on their topics, in order to generate a summary of the main topics for which we also offer a timelined view.

Fig. 4. Prototype application for the sector television

We group incoming news through document clustering. First we perform topic modeling using a bag-of-words representation with the vectors based on tf/idf values (Sect. 2.4). The clusters are fed to the multi-document summarisation service, summarising on a per-topic-basis and then to the timelining and multi-document summarisation service (fixed summary length of 200 words). Given that the same number of documents was clustered into topics using six different models (bag of words and tf/idf for HDP, LDA, and LSI each) and that the length of the summary for each topic was fixed at a maximum of 200 words, we discovered that the bag of words approach yields lengthier summaries than tf/idf. Additionally, in 90% of the cases, cluster-based link analysis outperformed all other approaches in terms of summary length.

The approach resembles Google News in that it focuses on entities whereas we want to move to an event-based representation of storylines. Recognised named entities focus on locations, persons or temporal expressions. To arrive at building blocks for storylines and also to identify modified news regarding stories already seen, we want to focus on the actual *events* mentioned in the document collection. We repeated the experiments by using the events extracted from the documents as features for clustering and applied the same algorithms (LSI, LDA, HDP) to the text associated with the extracted event. The clusters were summarised and timelined. While the individual services can be improved using more domain-specific training data, the goal is to present the processing results in a way that speeds up (human) access and understanding and that supports the journalist telling an interesting news story. In a previous experiment (Moreno-Schneider et al. 2016) we used biographical data and corresponding templates, which were presented as content pieces; in the present paper, events serve the same purpose. Figure 4 shows the current prototypes, i.e., entity and temporal expression annotation layers to visualise persons, locations and organisations, a timeline-based view and a map visualisation. The processing generates potential

storylines for the editor who can then use them to compose stories based on these automatically extracted key facts (collated from multiple documents). The GUI provides interactive browsing and exploring of the processed news collection. Figure 5 shows an example generated from our collection of local news data.

a) [. . .] the Federal Office for migration and refugees counts for this year, obviously with 600-thousand newly arriving asylum seekers [. . .]
b) [. . .] Potsdam is now the authority of 600-thousand newly arriving refugees – 100-thousand more so than predicted.
c) [. . .] the cabinet today announced the opportunities for young asylum seekers to work better.

Fig. 5. Example: Story on refugees, composed from 18 topics and 151 documents

4 Related Work

Semantic Storytelling can be defined as the generation of stories, identification of story paths or recommendation of storylines based on a certain set of content using a concrete narrative style or voice. Thus, automatic storytelling consists of two components: a semantic representation of story structure, and the ability to automatically visualise or generate a story from this semantic representation using some form of Natural Language Generation (NLG) (Rishes et al. 2013). In NLG, notable related work is described, among others, by (Jorge et al. 2013; Dionisio et al. 2016; Mazeika 2016; Farrell and Ware 2016). While an interesting discipline that is essential to applying any system aimed at automatically generating stories, especially regarding surface realisation, we primarily focus on the generation of the semantic structure of the story.

Bowden et al. (2016) describe what a story is and how to convert it into a dialogue story, i.e., a system capable of telling a story and then retelling it in different settings to different audiences. They define a story as a set of events, characters, and properties of the story, as well as relations among them, including reactions of characters to story events. For this they use EST (Rishes et al. 2013), a framework that produces a story annotated for the tool Scheherazade as a list of Deep Syntactic Structures, a dependency-tree structure where each node contains the lexical information for the important words in a sentence. Kybartas and Bidarra (2015) present GluNet, a flexible, open source knowledge-base that integrates a variety of lexical databases and facilitates commonsense reasoning for the definition of stories.

Similar to our approach is the work of Samuel et al. (2016). They describe a writing assistant that provides suggestions for the actions of characters. This assistant is meant to be a “playful tool”, which is intended to “serve the role of a digital writing partner”. We perform similar processes when extracting events and entities from a document collection but our system operates on a more general level and is meant to be applied in different professional sectors.

Several related approaches concentrate on specific domains. A few systems focus on providing content for entertainment purposes (Wood 2008), others

focus on storytelling in gaming (Gervás 2013), for recipes (Cimiano et al. 2013; Dale 1989) or weather reports (Belz 2008), requiring knowledge about characters, actions, locations, events, or objects that exist in this particular domain (Riedl and Young 2010; Turner 2014). A closely related approach is the one developed by Poulakos et al. (2015), which presents "an accessible graphical platform for content creators and even end users to create their own story worlds, populate it with smart characters and objects, and define narrative events that can be used by existing tools for automated narrative synthesis".

5 Conclusions

We developed curation technologies that can be applied in the sector-specific use cases of companies active in different sectors and content curation use cases. The partner companies are in need of semantic storytelling solutions that support their own in-house or their customers' content curators putting together new content products, either museum exhibitions, interactive online versions of public archives, news articles or news programmes. The motivation is to make the curators more efficient, to delegate routine tasks to the machine and to enable curators to produce higher quality products because the machine may be able to identify interesting, novel, eye-opening relationships between two pieces of content that a human is unable to recognise. The technologies, prototypically implemented and successfully applied in four sectors, show very promising results (Rehm et al. 2017a), even though the individual implementations of the interactive storytelling approaches are quite specific.

For the museums and exhibitions case we developed a prototype that allows the interactive curation, analysis and exploration of the background material of a new exhibition, supporting the knowledge workers who design the exhibition in their storytelling capabilities by helping them to identify interesting relationships. For the public archive case we implemented a prototype that semantically enriches a collection of letters so that a human expert can more efficiently tell interesting stories about the content – in our example we help the human curator to produce a travelogue about the different trips of the Mendelsohns as an alternative "view" upon the almost 2,800 letters. For newspaper journalism, we annotate named entities to generate clusters of documents that can be used as storylines. For the television case we applied a similar approach but we cluster events instead of named entities (including timelining).

This article provides a current snapshot of the technologies and approaches developed in our project. In a planned follow-up project we will experiment with Natural Language Generation approaches in order to produce natural language text – either complete documents or draft skeletons to be checked, revised and completed by human experts – based on automatically extracted information and on external knowledge provided as Linked Open Data. For this approach we anticipate a whole new set of challenges with regard to semantic storytelling.

Acknowledgments. The authors would like to thank the reviewers for their insightful comments and suggestions. The project "Digitale Kuratierungstechnologien" (DKT) is supported by the German Federal Ministry of Education and Research (BMBF), "Unternehmen Region", instrument Wachstumskern-Potenzial (no. 03WKP45). More information: http://www.digitale-kuratierung.de.

References

Belz, A.: Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. Nat. Lang. Eng. **14**(4), 431–455 (2008). https://doi.org/10.1017/S1351324907004664. ISSN 1351–3249

Bienert, A., de Wit, W., (eds.): EMA - Erich Mendelsohn Archiv. Der Briefwechsel von Erich und Luise Mendelsohn 1910–1953. Staatliche Museen zu Berlin, The Getty Research Institute, Los Angeles, March 2014. http://ema.smb.museum

Bourgonje, P., Moreno-Schneider, J., Nehring, J., Rehm, G., Sasaki, F., Srivastava, A.: Towards a platform for curation technologies: enriching text collections with a semantic-web layer. In: Sack, H., Rizzo, G., Steinmetz, N., Mladenić, D., Auer, S., Lange, C. (eds.) ESWC 2016. LNCS, vol. 9989, pp. 65–68. Springer, Cham (2016a). https://doi.org/10.1007/978-3-319-47602-5_14. ISBN 978-3-319-47602-5

Bourgonje, P., Moreno-Schneider, J., Rehm, G., Sasaki, F.: Processing document collections to automatically extract linked data: semantic storytelling technologies for smart curation workflows. In: Gangemi, A., Gardent, C., (eds.) Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016), Edinburgh, UK, September 2016b, pp. 13–16. The Association for Computational Linguistics (2016b)

Bowden, K.K., Lin, G.I., Reed, L.I., Fox Tree, J.E., Walker, M.A.: M2D: monolog to dialog generation for conversational story telling. In: Nack, F., Gordon, A.S. (eds.) ICIDS 2016. LNCS, vol. 10045, pp. 12–24. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48279-8_2. ISBN 978-3-319-48278-1

Cimiano, P., Lüker, J., Nagel, D., Unger, C.: Exploiting ontology lexica for generating natural language texts from RDF data. In: Proceedings of the 14th European Workshop on Natural Language Generation, Sofia, Bulgaria, August 2013, pp. 10–19. Association for Computational Linguistics (2013). http://www.aclweb.org/anthology/W13-2102

Dale, R.: Cooking up referring expressions. In: Proceedings of the 27th Annual Meeting on Association for Computational Linguistics, ACL 1989, Stroudsburg, PA, USA, pp. 68–75. Association for Computational Linguistics (1989). https://doi.org/10.3115/981623.981632

Dionisio, M., Nisi, V., Nunes, N., Bala, P.: Transmedia storytelling for exposing natural capital and promoting ecotourism. In: Nack, F., Gordon, A.S. (eds.) ICIDS 2016. LNCS, vol. 10045, pp. 351–362. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48279-8_31. ISBN 978-3-319-48279-8

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The automatic content extraction (ACE) program - tasks, data, and evaluation. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, May 2004. ELRA (2004)

Erkan, G., Radev, D.R.: LexPageRank: prestige in multi-document text summarization. In: EMNLP, Barcelona, Spain (2004)

Farrell, R., Ware, S.G.: Predicting user choices in interactive narratives using indexter's pairwise event salience hypothesis. In: Nack, F., Gordon, A.S. (eds.) ICIDS 2016. LNCS, vol. 10045, pp. 147–155. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48279-8_13. ISBN 978-3-319-48279-8

Gervás, P.: Stories from games: content and focalization selection in narrative composition. In: Proceedings of the I Spanish Symposium on Entertainment Computing, Universidad Complutense de Madrid, Madrid, Spain, September 2013

Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev. **53**(2), 217–288 (2011). https://doi.org/10.1137/090771806. ISSN 0036–1445

Hoffman, M., Bach, F.R., Blei, D.M.: Online learning for Latent Dirichlet Allocation. In: Lafferty, J.D., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (eds.) Advances in Neural Information Processing Systems 23, pp. 856–864. Curran Associates Inc. (2010). http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf

Jorge, C., Nisi, V., Nunes, N.J., Innella, G., Caldeira, M., Sousa, D.: Ambiguity in design: an airport split-flap display storytelling installation. In: Mackay, W.E., Brewster, S.A., Bødker, S., (eds.) 2013 ACM SIGCHI Conference on Human Factors in Computing Systems, CHI 2013, Paris, France, pp. 541–546. ACM (2013). https://doi.org/10.1145/2468356.2468452. ISBN 978-1-4503-1952-2

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Zens, R., Federico, M., Bertoldi, N., Dyer, C., Cowan, B., Shen, W., Moran, C., Bojar, O., Constantin, A., Herbst, E.: Moses: open source toolkit for statistical machine translation. In: Proceedings of ACL 2007, Prague, Czech Republic, pp. 177–180. ACL (2007)

Kybartas, B., Bidarra, R.: A semantic foundation for mixed-initiative computational storytelling. In: Schoenau-Fog, H., Bruni, L.E., Louchart, S., Baceviciute, S. (eds.) ICIDS 2015. LNCS, vol. 9445, pp. 162–169. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-27036-4_15. ISBN 978-3-319-27035-7

Mani, I., Maybury, M.T. (eds.): Advances in Automatic Text Summarization. MIT Press, Cambridge (1999)

Mazeika, J.: A rules-based system for adapting and transforming existing narratives. In: Nack, F., Gordon, A.S. (eds.) ICIDS 2016. LNCS, vol. 10045, pp. 176–183. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48279-8_16. ISBN 978-3-319-48279-8

McCallum, A.K.: MALLET: a machine learning for language toolkit (2002). http://mallet.cs.umass.edu

Moreno-Schneider, J., Bourgonje, P., Rehm, G.: Towards user interfaces for semantic storytelling. In: Yamamoto, S. (ed.) HIMI 2017, Part II. LNCS, vol. 10274, pp. 403–421. Springer, Cham (2017a). https://doi.org/10.1007/978-3-319-58524-6_32

Moreno-Schneider, J., Srivastava, A., Bourgonje, P., Wabnitz, D., Rehm, G.: Semantic storytelling, cross-lingual event detection and other semantic services for a newsroom content curation dashboard. In: Popescu, O., Strapparava, C. (eds.) Proceedings of the Second Workshop on Natural Language Processing meets Journalism - EMNLP 2017 Workshop (NLPMJ 2017), Copenhagen, Denmark, September 2017b, pp. 68–73 (2017b)

Moreno-Schneider, J., Bourgonje, P., Nehring, J., Rehm, G., Sasaki, F., Srivastava, A.: Towards semantic story telling with digital curation technologies. In: Birnbaum, L., Popescuk, O., Strapparava, C. (eds.) Proceedings of Natural Language Processing Meets Journalism - IJCAI-16 Workshop (NLPMJ 2016), New York, July 2016

Ou, S., Khoo, C.S.-G., Goh, D.H.: Design and development of a concept-based multi-document summarization system for research abstracts. J. Inf. Sci. **34**(3), 308–326 (2008). https://doi.org/10.1177/0165551507084630

Poulakos, S., Kapadia, M., Schüpfer, A., Zünd, F., Sumner, R., Gross, M.: Towards an accessible interface for story world building. In: AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, pp. 42–48 (2015). http://www.aaai.org/ocs/index.php/AIIDE/AIIDE15/paper/view/11583

Radev, D.R., Jing, H., Budzikowska, M.: Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In: Proceedings of the 2000 NAACL-ANLP-Workshop on Automatic Summarization, NAACL-ANLP-AutoSum 2000, Stroudsburg, PA, USA, pp. 21–30. ACL (2000). https://doi.org/10.3115/1117575.1117578

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., Manning, C.: A multi-pass sieve for coreference resolution. In: Proceedings of the 2010 Conference on Empirical Methods in NLP, EMNLP 2010, Stroudsburg, PA, USA, pp. 492–501. ACL (2010). http://dl.acm.org/citation.cfm?id=1870658.1870706

Rehm, G., He, J., Moreno-Schneider, J., Nehring, J., Quantz, J.: Designing user interfaces for curation technologies. In: Yamamoto, S. (ed.) HIMI 2017, Part I. LNCS, vol. 10273, pp. 388–406. Springer, Cham (2017a). https://doi.org/10.1007/978-3-319-58521-5_31

Rehm, G., Moreno-Schneider, J., Bourgonje, P., Srivastava, A., Nehring, J., Berger, A., König, L., Räuchle, S., Gerth, J.: Event detection and semantic storytelling: generating a travelogue from a large collection of personal letters. In: Caselli, T., Miller, B., van Erp, M., Vossen, P., Palmer, M., Hovy, E., Mitamura, T. (eds) Proceedings of the Events and Stories in the News Workshop, Vancouver, Canada, August 2017b, pp. 42–51. Association for Computational Linguistics. Co-located with ACL 2017 (2017b)

Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, May 2010, pp. 45–50. ELRA (2010). http://is.muni.cz/publication/884893/en

Riedl, M.O., Young, R.M.: Narrative planning: balancing plot and character. J. Artif. Int. Res. **39**(1), 217–268 (2010). http://dl.acm.org/citation.cfm?id=1946417.1946422. ISSN 1076–9757

Rishes, E., Lukin, S.M., Elson, D.K., Walker, M.A.: Generating different story tellings from semantic representations of narrative. In: Koenitz, H., Sezen, T.I., Ferri, G., Haahr, M., Sezen, D., Ç atak, G. (eds.) ICIDS 2013. LNCS, vol. 8230, pp. 192–204. Springer, Cham (2013). https://doi.org/10.1007/978-3-319-02756-2_24

Samuel, B., Mateas, M., Wardrip-Fruin, N.: The design of *Writing Buddy*: a mixed-initiative approach towards computational story collaboration. In: Nack, F., Gordon, A.S. (eds.) ICIDS 2016. LNCS, vol. 10045, pp. 388–396. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48279-8_34

Sasaki, F., Gornostay, T., Dojchinovski, M., Osella, M., Mannens, E., Stoitsis, G., Richie, P., Declerck, T., Koidl, K.: Introducing FREME: deploying linguistic linked data. In: Proceedings of the 4th Workshop of the Multilingual Semantic Web, MSW 2015 (2015)

DBPedia Spotlight. DBPedia Spotlight Website (2016). https://github.com/dbpedia-spotlight/

Srivastava, A., Sasaki, F., Bourgonje, P., Moreno-Schneider, J., Nehring, J., Rehm, G.: How to configure statistical machine translation with linked open data resources. In: Proceedings of Translating and the Computer 38 (TC38), pp. 138–148, London, UK, November 2016

Srivastava, A., Weber, S., Bourgonje, P., Rehm, G.: Different German and English coreference resolution models for multi-domain content curation scenarios. In: Rehm, G., Declerck, T. (eds.) GSCL 2017. LNAI, vol. 10713, pp. 48–61. Springer, Heidelberg (2017)

Strötgen, J., Gertz, M.: Multilingual and cross-domain temporal tagging. Lang. Resour. Eval. **47**(2), 269–298 (2013). https://doi.org/10.1007/s10579-012-9179-y

Turner, S.R.: The Creative Process: A Computer Model of Storytelling and Creativity. Taylor & Francis, Abingdon (2014). https://books.google.gr/books?id=1AjsAgAAQBAJ. ISBN 9781317780625

Wan, X., Yang, J.: Multi-document summarization using cluster-based link analysis. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in IR, SIGIR 2008, New York, NY, USA, pp. 299–306. ACM (2008). https://doi.org/10.1145/1390334.1390386. ISBN 978-1-60558-164-4

Wang, C., Paisley, J., Blei, D.M.: Online variational inference for the Hierarchical Dirichlet Process. In: Proceedings of the 14th International Conference on AI and Statistics (AISTATS), vol. 15, pp. 752–760 (2011). http://jmlr.csail.mit.edu/proceedings/papers/v15/wang11a/wang11a.pdf

Wick, M.: Geonames ontology (2015). http://www.geonames.org/about.html

Wood, M.D.: Exploiting semantics for personalized story creation. In: Proceedings of the International Conference on Semantic Computing, ICSC 2008, Washington, DC, USA, pp. 402–409. IEEE Computer Society (2008). https://doi.org/10.1109/ICSC.2008.10. ISBN 978-0-7695-3279-0

Yang, B., Mitchell, T.: Joint extraction of events and entities within a document context. In: Proceedings of the 2016 Conference of the North American Chapter of the ACL: Human Language Technologies, pp. 289–299. ACL (2016)