Semantic Storytelling, Cross-lingual Event Detection and other Semantic Services for a Newsroom Content Curation Dashboard

Julian Moreno-Schneider, Ankit Srivastava,
Peter Bourgonje, David Wabnitz*, Georg Rehm
DFKI GmbH, Language Technology Lab, Alt-Moabit 91c, 10559 Berlin, Germany
*Kreuzwerker GmbH, Ritterstraße 12-14, 10969 Berlin, Germany
Corresponding author: georg.rehm@dfki.de

Abstract

We present a prototypical content curation dashboard, to be used in the newsroom, and several of its underlying semantic content analysis components (such as named entity recognition, entity linking, summarisation and temporal expression analysis). The idea is to enable journalists (a) to process incoming content (agency reports, twitter feeds, reports, blog posts, social media etc.) and (b) to create new articles more easily and more efficiently. The prototype system also allows the automatic annotation of events in incoming content for the purpose of supporting journalists in identifying important, relevant or meaningful events and also to adapt the content currently in production accordingly in a semi-automatic way. One of our long-term goals is to support journalists building up entire storylines with automatic means. In the present prototype they are generated in a backend service using clustering methods that operate on the extracted events.

1 Introduction

Journalists write and distribute news articles based on information collected from different sources (news agencies, media streams, other news articles etc.). In order to produce a high-quality piece, a fair bit of research is needed on the topic and domain at hand. Facts have to be checked, requiring at least basic domain knowledge, different view points considered and information from multiple sources combined in a sensible way. In short, much research is needed to arrive at a piece of content that combines new and surprising information with a decent context of the event reported upon. While the amount of available, especially digital information is increasing on a daily basis, the journalist's ability to read all this information is decreasing in the little time available. This calls for tools that support journalists in processing large amounts of incoming information.

There are differences in journalistic reporting, depending on the event being covered. News about a event with global relevance, such as a war, differs from news about the inauguration of a local cultural centre. When looking at the available source material, the amount of background information, also its diversity, differs significantly in both cases. Coverage for the global event depends on a much larger amount of readily available information while the local event coverage depends on smaller amounts of data that may need a bit of effort in tracking them down (the name of local news sources for example). To address this difference in research requirements we describe a prototypical approach for cross-lingual semantic analysis, especially event detection, ultimately aimed at supporting journalists through semantic storytelling, based on two data sets.

Section 2 briefly describes the content curation dashboard, while Section 3 focuses upon semantic storytelling. Section 4 describes the use cases and sketches the results of initial experiments on news data. Section 5 concludes the paper.

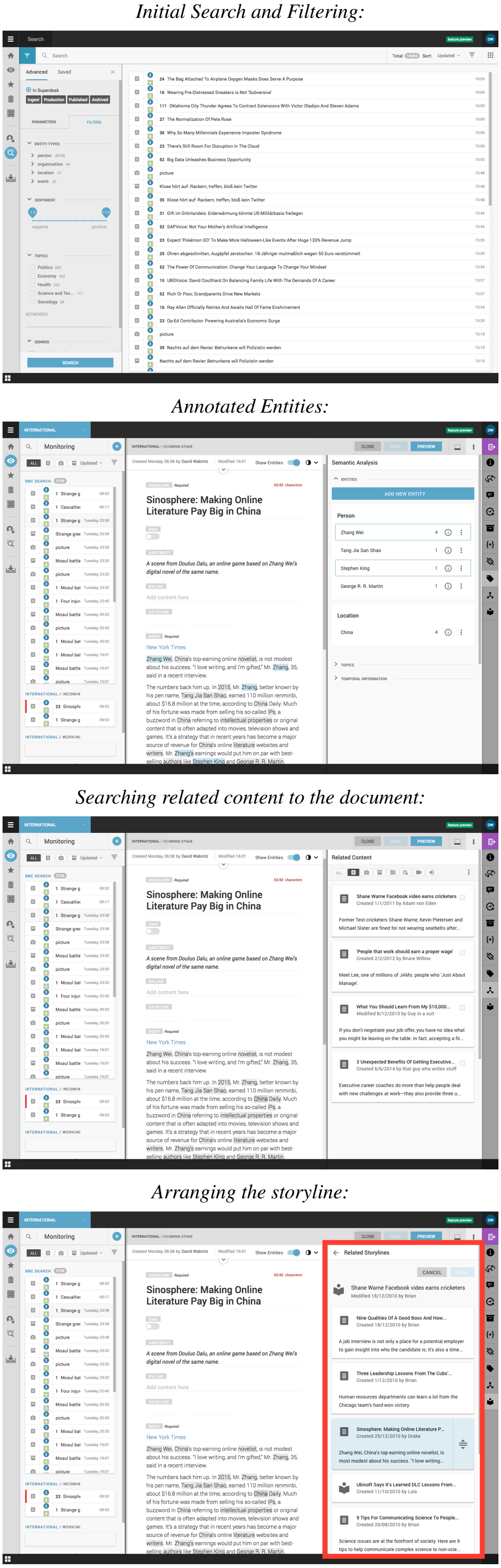
Proceedings of the 2017 EMNLP Workshop on Natural Language Processing meets Journalism, pages 68–73
Copenhagen, Denmark, September 7, 2017. ©2017 Association for Computational Linguistics

2 The Content Curation Dashboard

One of the partner companies involved in our joint technology transfer project (Rehm and Sasaki, 2015) to introduce curation technologies to different sectors, is currently designing and developing an extension for the open source newsroom software Superdesk.[1] Superdesk is a production environment for journalists that specialises on the creation of content, i. e., the play-out and rendering of the content is taken care of by other parts of a larger system. Our Superdesk extension allows the semantic processing of incoming news streams to enable several smart features, e. g., keyword alerts, content exploration, identifying related content, among others. The tool also allows the visualisation and annotation of news documents using additional information sources, databases and knowledge graphs such as Linked Data. The idea is to allow rich faceted search scenarios so that the journalist has fine-grained mechanisms for locating the needle in a potentially very large haystack of digital data. Faceted search includes entities, topics, sentiment values and genres, complemented with semantic information from external sources (DBpedia) enabling higher semantic search precision based on extracted information than would be possible with keyword based search.

Visualisation includes menus that show the annotated entities and their frequencies next to a set of related documents. Example screens of the content curation dashboard are shown in Figure 1. The Superdesk extension and the underlying semantic technologies mainly operate on the (1) ingest view and the (2) authoring view. The first view allows to ingest multiple incoming content channels into the production environment; our semantic tools can automatically analyse the content using, for example, named entity recognition, sentiment analysis, topic detection, classification (e. g., IPTC topics) and others, so that journalists can tailor the incoming news feed exactly to their liking and current topics. In the second view, the semantic tools are used to support the authoring process, to add and modify annotations, to recommend related content potentially to be linked to in the new article.

While Superdesk is a specialised newsroom software, journalists also often use Content Management Systems such as Symphony to automate day-to-day work (e. g., file and document man-



Initial Search and Filtering:

24 The Bag Attached To Airplane Oxygen Masks Does Serve A Purpose 16:05
16 Wearing Pre-Distressed Sneakers Is Not 'Subversive' 16:04
111 Oklahoma City Thunder Agrees To Contract Extensions With Victor Oladipo And Steven Adams 16:00
27 The Normalization Of Pete Rose 16:00
36 Why So Many Millennials Experience Imposter Syndrome 15:56
23 There's Still Room For Disruption In The Cloud 15:56
52 Big Data Unleashes Business Opportunity 15:53
picture 15:48
Klose hört auf: Rackern, treffen, bloß kein Twitter 15:48
35 Klose hört auf: Rackern, treffen, bloß kein Twitter 15:48
31 Gift im Grönlandeis: Erderwärmung könnte US-Militärbasis freilegen 15:45
52 SAPVoice: Not Your Mother's Artificial Intelligence 15:45
23 Expect 'Pokémon GO' To Make More Halloween-Like Events After Huge 133% Revenue Jump 15:35
25 Ohren abgeschnitten, Augäpfel zerstochen: 18-Jähriger mutmaßlich wegen 50 Euro verstümmelt 15:35
52 The Power Of Communication: Change Your Language To Change Your Mindset 15:34
15 UBSVoice: David Coulthard On Balancing Family Life With The Demands Of A Career 15:27
52 Rich Or Poor, Grandparents Drive New Markets 15:27
16 Ray Allen Officially Retires And Awaits Hall Of Fame Enshrinement 15:24
23 Op-Ed Contributor: Powering Australia's Economic Surge 15:20
picture 15:19
39 Nachts auf dem Revier: Betrunkene will Polizistin werden 15:19
Nachts auf dem Revier: Betrunkene will Polizistin werden 15:19

Annotated Entities:

Created Monday, 06:58 by David Wabnitz | Modified 16:01

HEADLINE Required 53/42 characters

Sinosphere: Making Online Literature Pay Big in China

ABSTRACT

A scene from Douluo Dalu, an online game based on Zhang Wei's digital novel of the same name.

BYLINE Add content here

BODY Required

New York Times

Zhang Wei, China's top-earning online novelist, is not modest about his success. "I love writing, and I'm gifted," Mr. Zhang, 35, said in a recent interview.

The numbers back him up. In 2015, Mr. Zhang, better known by his pen name, Tang Jia San Shao, earned 110 million renminbi, about $16.8 million at the time, according to China Daily. Much of his fortune was made from selling his so-called IPs, a buzzword in China referring to intellectual properties or original content that is often adapted into movies, television shows and games. It's a strategy that in recent years has become a major source of revenue for China's online literature websites and writers. Mr. Zhang's earnings would put him on par with best-selling authors like Stephen King and George R. R. Martin.

Semantic Analysis

Person	
Zhang Wei	4
Tang Jia San Shao	1
Stephen King	1
George R. R. Martin	1
Location	
China	4

Searching related content to the document:

Related Content

Shane Warne Facebook video earns cricketers
Created 1/1/2011 by Adam von Eden
Former Test cricketers Shane Warne, Kevin Pietersen and Michael Slater are fined for not wearing seatbelts after...

'People that work should earn a proper wage'
Created 2/2/2012 by Bruce Willow
Meet Lee, one of millions of JAMs: people who 'Just About Manage'.

What You Should Learn From My $10,000...
Modified 8/12/2015 by Guy in a suit
If you don't negotiate your job offer, you have no idea what you might be leaving on the table. In fact, accepting a fir...

3 Unexpected Benefits Of Getting Executive...
Created 6/6/2014 by that guy who writes stuff
Executive career coaches do more than help people deal with new challenges at work—they also provide three u...

Arranging the storyline:

Related Storylines

Shane Warne Facebook video earns cricketers
Modified 18/12/2010 by Brian

Nine Qualities Of A Good Boss And How...
Created 18/12/2010 by Brian
A job interview is not only a place for a potential employer to gain insight into who the candidate is; it's also a time...

Three Leadership Lessons From The Cubs'...
Created 1/12/2010 by Brian
Human resources departments can learn a lot from the Chicago team's hard-won victory.

Sinosphere: Making Online Literature P...
Created 29/12/2010 by Drake
Zhang Wei, China's top-earning online novelist, is most modest about his success. "I love writing...

Ubisoft Says It's Learned DLC Lessons From...
Created 11/10/2010 by Lois

9 Tips For Communicating Science To People...
Created 20/08/2010 by Brian
Science issues are at the forefront of society. Here are 9 tips to help communicate complex science to non-scie...

Figure 1: The Content Curation Dashboard

[1] https://www.superdesk.org

agement).[2] For document exploration proper IR systems (mainly ElasticSearch) have gained popularity through user-friendly wrappers such as Kibana,[3] which offers visualisation of maps and timelines, or Kibi,[4] which offers richer visualisation capabilities (graphs, charts).

We want to enable journalists interactively to put together a story based on extensive semantic content enrichment. In our various use cases, different parts of the content function as atomic building blocks (sentences, paragraphs, documents). In the use case of the present paper we focus, for now, upon document level building blocks for generating stories, i. e., documents can be rearranged, included and deleted from a storyline. In a later stage we plan to use smaller content components with which we will experiment towards the generation of news articles based on multiple story paths, automatically generated with the help of semantic annotations.

Currently our technology allows us to generate "semantic fingerprints" which enable suggestions as to where the currently edited article would fit in either an existing storyline or recommend a new one based on content currently being readied for production and further enhanced by related content. Generally, it would enable journalists to work more in smaller update increments to a developing story or storyline without having to retell a story time and again. Storylines, in that sense, can be thought of as a container, pulling related stories together. By providing for a semi-automatic arrangement of stories within a storyline (e. g., by suggesting that a story fits in a certain slot within a chronological order of related events or in a certain slot in an existing narrative), journalists can be alleviated of the need to be aware of these relationships and to manage them. Consumers of news articles get the benefit of additional context and navigational tools provided by articles being arranged in storylines as well as enjoying a more efficient news consumption experience.

3 Semantic Storytelling for Journalists

In our current prototype we attempt to generate a set of potential storylines from a collection of incoming news documents. A storyline is a set of building blocks that, through their combination (temporal, geographical, semantic, etc.) form a story. In several use cases, the atomic building blocks are documents; we use a more fine-grained approach in which events are the building blocks of storylines, i. e., a set of entities governed by a trigger element (normally a verb) and which together represent an occurring action in the text.

The linguistic processing is done by a semantic annotation pipeline that creates a layer of named entities, temporal expressions and other information on top of the document collection, also augmented with event detection (Bourgonje et al., 2016a,b). For example, in the sentence "Barack Obama visited Ukraine last summer and had a meeting with Petró Poroshenko" there are two persons (Barack Obama, Petró Poroshenko), one location (Ukraine) and one temporal expression (last summer). There are also two trigger verbs: "visited" and "meet". Therefore, there are two events in this sentence: [visited, Barack Obama, Ukraine, last summer] and [meet, Barack Obama, Petró Poroshenko]. The sentence "Vladimir Putin will meet next summer with president Petró Poroshenko in Moscow" contains one event: [meet, Vladimir Putin, Petró Poroshenko, Moscow, next summer]. Events in German or English texts are extracted using cross-lingual event detection (Section 3.1). Then, storylines are created from combinations of annotated events using three clustering techniques to obtain related and similar information in the collection. In the Superdesk extension (Section 2), storylines are still composed of a set of related documents. Once completed, the extension will also operate on the more fine-grained event-centric approach.

Work related to our rather general, domain-independent storytelling concept typically concentrates on very specific objectives. A few systems focus on providing content for entertainment purposes (Wood, 2008; Poulakos et al., 2015). Other researchers focus on specific domains, e. g., storytelling in gaming (Gervás, 2013), for recipes (Cimiano et al., 2013)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). Regarding news, (Shahaf and Guestrin, 2010) describe methods for navigating within a new topic using hidden connections. They automatically find links between news articles through the extraction of links between entities. (Vossen

[2] https://symphony.com

[3] https://www.elastic.co/products/kibana

[4] https://siren.solutions/kibi/

et al., 2015) handle news streams through a formal model for representing storylines. They also describe a first implementation and visualisation that helps inspecting the generated structures.

3.1 Cross-lingual Event Detection

We implemented a cross-lingual event detection system, i. e., we automatically translate non-English documents to English using Moses (Koehn et al., 2007) and detect events in the translated documents using a state-of-the-art event extraction system based on (Yang and Mitchell, 2016), trained on the English section of ACE 2005 (Doddington et al., 2004). The cross-lingual detection of events encompasses a pipeline that ends up with a list of annotated events in every document (Rehm et al., 2017b).

3.2 Semantic Storytelling

Extracted events themselves are not useful to a journalist who works on a set of documents. They have to be analysed further, summarised, rearranged and then presented in a way that speeds up (human) access and understanding. In a previous approach (Schneider et al., 2016) we focused upon template-filling, using the results of relation extraction to fill (biography) templates to present these as content pieces to the knowledge worker. In the present paper, events serve the same purpose, delivering content pieces for a news article.

These general clusters of events can provide the logical text structure of a new journalistic piece. We can also cluster documents based on the temporal dimension grouping together events that happened in the same period of time (e. g., a war or an award ceremony), or based on locations, using latitude and longitude coordinates. Another option is traditional semantic clustering, obtaining sets of documents that talk about the same events and entities. To get semantically related events, we cluster documents based on the entities that appear in the events (entity frequency). Our interpretation of semantically related events are events that share entities as their participants (subject and object). The cluster information for the two previous examples is shown in Table 1. Once the documents are clustered, their events are grouped and ranked by frequency. In the previous example there were three events, one document with two events and one document with one event: $d1 = \{ev1, ev2\}$ and $d2 = \{ev3\}$. If both documents are in the same cluster, and considering that we use the clusters as storylines, the resulting set of events in this storyline will be $\{ev1, ev2, ev3\}$.

3.3 Visualisation

To get a better understanding of the data set, the analysis results, the extracted events and to prepare attaching the semantic storytelling back-end to the newsroom content curation dashboard, we implemented an experimental visualisation prototype (Figure 2). This prototype provides access to the full set of semantic analysis information and can be used interactively to explore and evaluate the system. The map shows locations involved in extracted events with highlighted annotations. The slider below the map can be used to filter events by time. Additional details and case studies can be found in (Rehm et al., 2017a; Schneider et al., 2017). We will explore if we can integrate part of this prototype tool into the Superdesk extension.

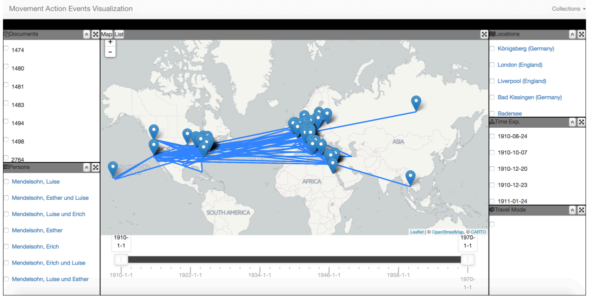


Figure 2: The experimental visualisation tool

4 Evaluation

We performed a qualitative evaluation of several generated storylines (clusters).[5] We apply the storytelling generation approach in two journalistic use cases: domain-specific web-crawled global news and general domain regional news. While the basic steps in both cases are the same (collecting relevant information, checking facts, writing an article, etc.), there are differences that make these two cases special: the “global news” articles are in English and were collected online while the “regional news” articles are in German, distributed by a news agency, so language usage and also register/style is different. We applied several clustering algorithms to both data sets using fixed and free cluster sizes; EM provides a rather balanced distribution along clusters.

[5] We would have performed the evaluation with the Superdesk extension but are still in the process of fully integrating the current prototype.

	Obama	Petró	Summer	Putin	Moscow	Ukraine
Document $d1$	1	1	1	0	0	1
Document $d2$	0	1	1	1	1	0

Table 1: Using the frequencies of extracted events as features for a clustering algorithm

4.1 Global News: Obama's Trips

For the global news case we used a data set that consists of news articles on the trips of Barack Obama (487 files, 24,387 sentences, 897,630 tokens).[6] All documents are English online news, which is why boilerplate cleaning was applied.[7]. The storytelling backend annotated a total of 61,718 entity mentions and 6,752 event triggers. After clustering using EM and the 50 most frequent entities as features, we obtain five clusters, i. e., five storylines. The number of documents contained in each storyline is: 4, 4, 19, 19, 16. The number of events included in each cluster is: 472, 1027, 2525, 3785, 2638. In the first cluster, there are three documents talking about trips to Asia (China, Vietnam, Istanbul) and only one to Germany; the documents in the second cluster are grouped together because of remaining HTML tags after the boilerplate cleaning; the third cluster is semantically more diverse and contains documents talking about trips to South America (Cuba, Argentina, Colombia) and Asia (Beijing, Bangkok, Philippines), but there are some unrelated documents talking about wars (Syria, Second World War) and trips to the UK (London); the fourth cluster contains mainly documents talking about trips to Europe to the G20 Summit; the last cluster groups documents mentioning different places (Brazil, Egypt, Afghanistan, Japan, etc). Considering the topic diversity, this approach seems to be an interesting initial point for the automatic generation of semantic content for stories in large collections with topic heterogeneity.

4.2 Regional News

For the second use case we analyse a German general domain regional news collection (1,037 articles, 716,885 words, avg. number of words 691.3), provided by one of our project partner companies. The storytelling system, working on the automatically translated English documents, annotated a total of 61,054 entity mentions and 2,571 event triggers. The discrepancy in the number of events between the two data sets can be attributed to different writing styles as well as the fact that the latter was translated automatically. After the clustering process using EM and the 50 most frequent entities as features, we obtain five storylines (with 34, 17, 113, 25, 35 documents, and 4167, 2529, 11885, 2930, 3284 events, respectively). After manually evaluating the documents and events we can summarise that the automatic translation of the documents, performed with an MT system that had not been domain-adapted, has had a negative impact on the performance of the event extraction system and, therefore, the clustering results.

5 Conclusions and Future Work

We present a system based on three main components: (1) a cross-lingual event detection module; (2) a storyline generation component that can determine related events; (3) a newsroom content curation dashboard prototype that helps journalists in the process of analysing large document collections. Regarding the manual evaluation of the generated storylines, we observe that the storyline generator clearly unveils inherent semantic relatedness as a basic property of the documents in the global news data set, while demonstrating documents in the local news data set to be rather unrelated. Further improvement of the storyline generation and event detection system is foreseen for future work, especially regarding deeper and more fine-grained filtering of the extracted events in order to minimise the number of events included in a storyline. A future version of the newsroom curation dashboard will be able to suggest, to the journalist, event-based storylines. We will also include additional visualisation, as well as more linked data sources. In the semantic backend, additional processing modules will be included, especially coreference resolution to improve the coverage of extracted entity mentions.

[6] Based on a list of links to news articles in https://en.wikipedia.org/wiki/List_of_international_presidential_trips_made_by_Barack_Obama

[7] https://github.com/kohlschutter/boilerpipe

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful feedback. The project "Digitale Kuratierungstechnologien" is supported by the German Federal Ministry of Education and Research, Wachstumskern-Potenzial (no. 03WKP45). http://www.digitale-kuratierung.de.

References

Anja Belz. 2008. Automatic Generation of Weather Forecast Texts Using Comprehensive Probabilistic Generation-space Models. *Natural Language Engineering* 14(4):431–455.

Peter Bourgonje, Julian Moreno-Schneider, Jan Nehring, Georg Rehm, Felix Sasaki, and Ankit Srivastava. 2016a. Towards a Platform for Curation Technologies: Enriching Text Collections with a Semantic-Web Layer. In H. Sack, G. Rizzo, N. Steinmetz, D. Mladenic, S. Auer, and C. Lange, editors, *The Semantic Web*. Springer, number 9989 in LNCS, pages 65–68.

Peter Bourgonje, Julian Moreno Schneider, Georg Rehm, and Felix Sasaki. 2016b. Processing Document Collections to Automatically Extract Linked Data: Semantic Storytelling Technologies for Smart Curation Workflows. In A. Gangemi and C. Gardent, editors, *Proc. of the 2nd Int. Workshop on NLG and the Semantic Web (WebNLG 2016)*. ACL, Edinburgh, UK, pages 13–16.

Philipp Cimiano, Janna Lüker, David Nagel, and Christina Unger. 2013. Exploiting Ontology Lexica for Generating Natural Language Texts from RDF Data. In *Proc. of the 14th European Workshop on NLG*. ACL, Sofia, Bulgaria, pages 10–19.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ace) program – tasks, data, and evaluation. In *Proc. of LREC 2004*. ELRA, Lisbon, Portugal.

Pablo Gervás. 2013. Stories from Games: Content and Focalization Selection in Narrative Composition. In *I Spanish Symposium on Entertainment Computing*. Universidad Complutense de Madrid, Spain.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, R. Zens, M. Federico, N. Bertoldi, C. Dyer, B. Cowan, W. Shen, C. Moran, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. ACL, Prague, CZ, pages 177–180.

Steven Poulakos, Mubbasir Kapadia, Andrea Schüpfer, Fabio Zünd, Robert Sumner, and Markus Gross. 2015. Towards an Accessible Interface for Story World Building. In *AAAI Conf. on AI and Interactive Digital Entertainment*. pages 42–48.

Georg Rehm, Jing He, Julian Moreno Schneider, Jan Nehring, and Joachim Quantz. 2017a. Designing User Interfaces for Curation Technologies. In S. Yamamoto, editor, *Human Interface and the Management of Information: Information, Knowledge and Interaction Design, 19th Int. Conf., HCI Int. 2017*. Springer, Vancouver, CA, number 10273 in LNCS, pages 388–406. Part I.

Georg Rehm and Felix Sasaki. 2015. Digitale Kuratierungstechnologien – Verfahren für die effiziente Verarbeitung, Erstellung und Verteilung qualitativ hochwertiger Medieninhalte. In *Proc. of GSCL 2015*. pages 138–139.

Georg Rehm, Julian Moreno Schneider, Peter Bourgonje, Ankit Srivastava, Jan Nehring, Armin Berger, Luca König, Sören Räuchle, and Jens Gerth. 2017b. Event Detection and Semantic Storytelling: Generating a Travelogue from a large Collection of Personal Letters. In T. Caselli, B. Miller, M. van Erp, P. Vossen, M. Palmer, E. Hovy, and T. Mitamura, editors, *Proc. of the Events and Stories in the News Workshop*. ACL, Vancouver, CA.

Mark Owen Riedl and Robert Michael Young. 2010. Narrative Planning: Balancing Plot and Character. *J. Artif. Int. Res.* 39(1):217–268.

Julian Moreno Schneider, Peter Bourgonje, Jan Nehring, Georg Rehm, Felix Sasaki, and Ankit Srivastava. 2016. Towards Semantic Story Telling with Digital Curation Technologies. In L. Birnbaum, O. Popescu, and C. Strapparava, editors, *Proc. of NLP meets Journalism (NLPMJ 2016)*. NY.

Julian Moreno Schneider, Peter Bourgonje, and Georg Rehm. 2017. Towards User Interfaces for Semantic Storytelling. In S. Yamamoto, editor, *Human Interface and the Management of Information: Information, Knowledge and Interaction Design, 19th Int. Conf, HCI Int. 2017*. Springer, Vancouver, CA, number 10274 in LNCS, pages 403–421. Part II.

Dafna Shahaf and Carlos Guestrin. 2010. Connecting the dots between news articles. In *Proc. of 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, KDD '10, pages 623–632.

S.R. Turner. 2014. *The Creative Process: A Computer Model of Storytelling and Creativity*. Taylor & Francis.

Piek Vossen, Tommaso Caselli, and Yiota Kontzopoulou. 2015. Storylines for Structuring Massive Streams of News. In *Proc. of 1st Workshop on Computing News StoryLines (CNewS 2015), co-located with ACL 2015 and ACL-IJCNLP 2015*. Bejing, China.

Mark D. Wood. 2008. Exploiting Semantics for Personalized Story Creation. In *Proc. of Int. Conf. on Semantic Computing*. Washington, DC, ICSC '08, pages 402–409.

Bishan Yang and Tom Mitchell. 2016. Joint extraction of events and entities within a document context. In *Proc. of NAACL 2016*. ACL, pages 289–299.