Towards Layered Events and Schema Representations in Long Documents

Hans Ole Hatzel
Language Technology Group
Universität Hamburg, Germany
hatzel@informatik.uni-hamburg.de

Chris Biemann
Language Technology Group
Universität Hamburg, Germany
biemann@informatik.uni-hamburg.de

Abstract

In this thesis proposal, we explore event extraction and event representation on literary texts. Due to its variety of genres and varying document length, literature is a challenging domain, yet the representation of literary content has received relatively little attention. As most individual events contribute little to the overall semantics of literary documents, we model events at different granularities. On the conceptual level, we adapt the previous definition of schemas as sequences of events, all describing a single process connected through shared participants, and extend the notion to allow modeling a document's content using sequences of schemas. Technically, the segmentation of event sequences into schemas is approached by modeling such sequences, making use of the narrative cloze task, which is the prediction of masked events in event sequence contexts. We propose building on sequences of event embeddings to form schema representations, thereby summarizing sections of documents using a fixed-size representation. This approach will give rise to comparisons of sections such as chapters up to the comparison of entire literary works on the level of their schema structure, paving the way to a computational approach to quantitative literary research.

1 Introduction

Events generally describe any change of state (Hogenboom et al., 2016) and are often used in information extraction scenarios (Gaizauskas and Wilks, 1998; Niklaus et al., 2018). The modeling of sequences of events has the potential of aiding literary scientists in understanding narrative patterns and devices. Determining which events in a narrative are crucial is challenging and relates to a variety of related tasks, such as summarization, comparison, or even story generation. Understanding the contexts of an event requires modeling its arguments and semantics. A simple representation can be the subject and object relating to a given verb, in conjunction with the verb's lemma (Chambers and Jurafsky, 2008).

If one only wants to include events involving a single character in a story, it is necessary to consider only those predicates with arguments coreferring to the character. The narrative coherence assumption says that "verbs sharing coreferring arguments are semantically connected by virtue of narrative discourse structure" (Chambers and Jurafsky, 2008). Verbs connected in this way are, under the assumption, considered to be part of the same so-called narrative chain (Chambers and Jurafsky, 2008). Previous work has focused on finding chains as representations of narratives in short documents, combining individual narrative chains, each focused on one character, into a schema involving multiple chains and thereby multiple characters (Chambers and Jurafsky, 2009). While the overall narrative in a long document could be regarded as a large schema, a variety of sub-schemas exists describing each scene using individual events. As a result, a typical document in our domain contains multiple schemas.

Figure 1 illustrates a potential separation of an event sequence into schemas. For each event E_n^C in any given text we know, based on coreference resolution, which entities C are involved with it (i.e.: occur as its arguments). Intuitively a separation boundary is preferably found between non-connected events. The verbs *"leaving"* and *"arriving"*, for example, are strongly connected events; we expect them to often appear in sequence. After modeling the likelihood of different events occurring in sequence, we can calculate the model's perplexity with regard to a specific event and use this information for the separation of chains. Even in our simple example (Fig. 1) it is not clear where exactly to place separations, E_7 could, for example, form a *social gathering* schema with E_5 and E_6 instead of a separate *transportation* schema.
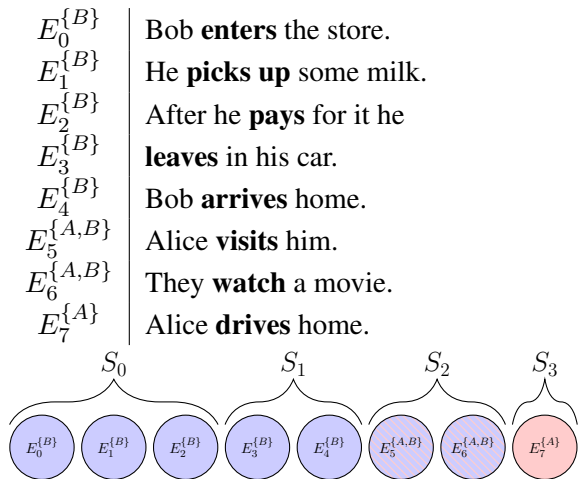
32

$E_0^{\{B\}}$	Bob **enters** the store.
$E_1^{\{B\}}$	He **picks up** some milk.
$E_2^{\{B\}}$	After he **pays** for it he
$E_3^{\{B\}}$	**leaves** in his car.
$E_4^{\{B\}}$	Bob **arrives** home.
$E_5^{\{A,B\}}$	Alice **visits** him.
$E_6^{\{A,B\}}$	They **watch** a movie.
$E_7^{\{A\}}$	Alice **drives** home.

Figure 1: One possible separation of the events into four schemas splits the event up into a *shopping*, a *transportation*, a *social gathering*, and another *transportation* schema.

2 Related Work

2.1 Event Processing

The detection of events has mostly focused on domains outside of literature, such as news (Doddington et al., 2004; Chambers and Jurafsky, 2008). More recently, Sims et al. (2019) created a new dataset of annotated literary texts.

2.2 Semantic Frame Induction

Semantic frames, in the context of FrameNet (Baker et al., 1998), are definitions of word senses where each sense can be evoked by multiple different words. The "Commerce_buy" frame, for example, can be evoked by the verbs "buy", "aquire" and "purchase", among others. FrameNet is an annotated dataset, marking for each predicate the frame that it evokes. A German frame resource called SALSA (Burchardt et al., 2006) builds on the frame lexicon provided by FrameNet.

The induction of specific frames has received much attention (Gildea and Jurafsky, 2000; Das et al., 2014). Generally, frame-semantic parsing is split into two sub-tasks of relevance to us: **(i)** target detection, the discovery of predicates evoking frames, and **(ii)** frame induction, the classification tasks of deciding which frame a predicate evokes (Das et al., 2014, p. 19). For the SemEval-2007 shared task (Pradhan et al., 2007), the work by Johansson and Nugues (2007) relies on the FrameNet lexicon specifying all possible frames for a predicate, with their model only deciding between the defined options. To handle predicates not covered

by FrameNet but occurring in the evaluation data they map uncovered verbs to existing ones using WordNet (Fellbaum, 1998).

Our proposal is closely related to QasemiZadeh et al. (2019), who introduce a shared task for unsupervised frame induction. Unlike the FrameNet dataset, they only provide frame annotations for verbs.

2.3 Event Sequences

Chambers and Jurafsky (2008) worked on learning narrative chains, sequences of events sharing a common protagonist. They operate on news data, introducing the narrative cloze task (the task of, given its surrounding events, predicting an event in a narrative chain). Chambers and Jurafsky (2009) extend the concept of narrative chains to narrative schemas, which involve more than one character and capture the interactions of different chains. Our approach is an extension of this work in that we aim to extract multiple schemas from a single long document. We assume that a document contains the descriptions of multiple processes or scenarios where each forms a schema.

Distinguishing real from generated event chains has been used in discriminative setups for story generation. Goldfarb-Tarrant et al. (2020) use event sequences as a building block to allow language models to generate globally consistent stories based on short prompts. Their model is trained to discern shuffled event sequences (using different shuffling strategies) from real ones. Guan et al. (2020) generate common-sense stories based on external knowledge bases. To our knowledge, no existing event modeling literature operates on longer chains of events as found in the domain of long-form literature.

Our approach is closely related to the one by Chambers and Jurafsky (2008) and Chambers and Jurafsky (2009), extending their approach to use vector representations over verb forms and to the operation on longer texts with multiple schemas.

2.4 Coreference Resolution

Coreference resolution is the task of identifying spans of text referring to the same entity within a document. Spans of text that refer to an entity are called mentions, in the sentence "[Alice] got up to greet [her] friend.", for example, both "Alice" and "her" refer to the same entity. The output of a coreference system is a set of mentions for each entity in the text. With the recent success of

contextual embedding based coreference resolution approaches (Xu and Choi, 2020; Joshi et al., 2019, 2020) and its adaptation to longer documents on English data (Xia et al., 2020; Toshniwal et al., 2020), it seems possible that learning-based approaches could outperform rule-based ones, even on documents the length of entire novels. For English, the CoNLL-2012 shared task, based on the OntoNotes 5.0 dataset, is universally used for evaluation (Pradhan et al., 2012). The improvement in performance on this task in the recent past has largely been attributed to the improvements in underlying embeddings (Xu and Choi, 2020). Existing approaches on German news-domain data (Roesiger and Kuhn, 2016) are based on rule-based systems.

LitBank (Bamman et al., 2020) is a dataset of English novels with coreference annotations. Recent approaches by Xia et al. (2020) and Toshniwal et al. (2020) have evaluated their approach on this dataset. Krug et al. (2015) have approached the domain of German literature using rule-based coreference resolution. They point out issues with machine learning approaches, namely the fact that literary text is very different from the news data usually used for training, and provide a corpus for evaluation (Krug et al., 2018). The availability and quality of pre-trained embeddings as well as the absence of very large annotated German literary datasets are hindrances to applying state-of-the-art English approaches. Recently neural networks, however, have been found to perform similarly to rule-based approaches in our domain, with weaknesses in global consistency (Krug, 2020, chap. 8).

3 Research Questions

Generally, the proposed thesis seeks to model broader narratives by building up from single events. We aim to build two-layered models, building from events to schemas by segmenting chains of events into semantically related sub-chains. Those sub-chains sharing coreferring arguments form what we call a schema. Over a simple sequence model of events, this has the potential benefits of allowing for human analysis and simplifying comparisons between multiple texts.

RQ 1: How can events be represented? We approach the detection of events by processing verb occurrences. We aim to make use of dense vector representations of frames instead of using discrete frames. This is motivated by coverage concerns as well as the intuitive insight that frames have vary-

ing semantic distances between each other, which we hope can be represented by vector space distances. The approach will be evaluated on existing semantic frame resources as well as regarding their contribution towards schemas.

RQ 2: How can schemas be represented? Through the use of sequence models, we will attempt to find semantically related sequences of events. This may mean finding common-sense event sequences. For example *"take cart" – "take fruit" – "queue up" – "pay"* clearly is a sequence of events typical for grocery shopping, even though no individual event is uniquely indicative of grocery shopping. In this way, we may find semantic structures in texts that only emerge from the combination of several events.

We will experiment with different approaches to transforming sequences of events into schema representations. A simple approach may be averaging of event representations; more advanced approaches involving neural sequence models are also to be explored.

RQ 3: Which role does coreference play in schema representations? Coreference allows us to resolve the arguments of frames to their entities. Predicates that share corefering arguments may, depending on the segmentation, be part of the same schema. Entities will be chosen based on their prevalence, only entities with multiple occurrences are of interest. As a result, all predicates not involved with entities of interest are discarded immediately; this is an implied filtering step removing many predicates that do not constitute events. Descriptions of scenery for example would usually be discarded in such a scenario. The evaluation of coreference resolution can be performed on existing datasets. Literary datasets generally only annotate characters, rather than all entities.

It is conceivable that representation learning on events in text order may, in our case, be an appropriate replacement for coreference resolution. In this case, the presence of multiple events in proximity would be modeled rather than an explicit interaction. Initial filtering of non-event predicates is, in this case, required to not include predicates irrelevant to the story at large.

RQ 4: How can event and schema representations be adapted to literary works. We hypothesize to encounter the following challenges in our approach to literary works: document length, vocabulary mismatch with pre-trained models, and

a diversity of domains (i.e.: different literary genres). To address these, we will explore the role of segmentation for processing documents in sections, the viability of incremental processing, and the role of pre-training and unsupervised fine-tuning. Aside from intrinsic evaluations of schemas based on their similarity and predicates based on them constituting events, we plan to derive summaries from the schema structure and compare them to human-generated summaries in literary lexicons (e.g. Arnold, 2009).

4 Methodology

From the research questions, two immediate directions emerge: event extraction, including coreference, and event representations. Later in the research process, we plan to build two-layer models transforming sequences of events into schemas.

4.1 Datasets

We operate on historical German literature in the form of the d-Prose dataset (Gius et al., 2020). Event annotations will, in cooperation with literary scientists, be created on a small subset of this data. In this subset, all verbs will be annotated, indicated whether or not they represent an event. For any verb that does represent an event, a set of binary features will be recorded, indicating several binary features based on concepts from narratology (Schmid, 2014). These features capture such criteria as reversibility, unexpectedness, and relevance of events.

4.2 Frame Identification for Event Representation

Initially, we assume each verb to evoke a frame and to represent an event, thereby addressing target detection using a parser-based heuristic. One notable exception, to the assumption of all verbs evoking frames, is stative verbs, *"Water is cold"* does not describe an event. Other cases such as inductive generalizations like *"Metal expands in the heat"* are more difficult to handle and may require machine learning approaches. Our initial approaches will only rely on the text order of events; we choose not to apply temporal ordering approaches (Mirroshandel et al., 2009; Mostafazadeh et al., 2016).

Concerns over insufficient coverage in the frame annotation data are motivated by an assumed diverse vocabulary in the domain of German literature. We separate coverage issues with frame re-

sources into two categories, expecting both to occur with our data: **(i)** missing frames where, as pointed out by Yong and Torrent (2020), some semantics may not be covered, and **(ii)** missing lexical units where not-before-seen verbs evoke known frames.

While previous work by Yong and Torrent (2020) addressed missing frame coverage concerns by generating new frames, our approach does not necessitate discrete frame representations, rather we see multiple potential benefits to using continuous representations instead. Vector representations for different frames may model their semantic distances, different frames of communication such as *"Statement"* and *"Reporting"*, for example, are relatively closely related. Further, continuous representations may cover gradual distinctions between frames. The lexical unit *"say"* will typically evoke the *"Statement"* frame, while the verb *"scream"* will evoke the *"Communication_noise"* frame; gradual decisions could be made as to which frame the example *"she spoke loudly"* should evoke. Lastly, continuous representations are a good fit for processing neural models, no additional embedding layer is needed.

Our initial approach mirrors the one described as "Bottom-up Prototype" by Sikos and Padó (2019). In this approach, for each frame, the average vector representation of all training examples is computed, with the resulting centroid representing the entire frame. With this approach, using BERT-based embeddings, assigning frames based on the closest centroid embedding, (Devlin et al., 2019) we only barely reached double-digit results (in terms of frame classification F1-score) without lexical unit filtering while predicting German SALSA frames. These current results are not comparable with existing ones that we are aware of but we will make sure to apply our approach to existing datasets (e.g. Pradhan et al., 2007) in the future to facilitate comparisons. To retain the wider applicability of our embeddings, while improving results, we decided to use an approach similar to the "Bottom-up plus Top-down Prototype" one taken by QasemiZadeh et al. (2019). We train a BERT network to decide if a pair of lexical units in their contexts evoke the same frame. Unlike QasemiZadeh et al. (2019), we rely on embedding similarity to frame centroids at evaluation time.

Model Name	Max F1-Score
bert-german	71.73
bert-dbmdz	72.71
multilingual-bert	73.36
bert-electra	**75.86**
IMS HotCoref DE[1]	48.54

Table 1: Preliminary F1 scores for German coreference resolution on the TüBa-D/Z 10 validation set for different underlying embeddings using early stopping, with previous results listed for comparison.

4.3 Coreference Resolution

Coreference resolution is required to extract chains of events sharing a specific entity. Our initial results are promising, showing that current neural approaches using modern embeddings perform very well on German data.

In the experiments we present in this proposal, we train and evaluate German coreference models on the TüBa-D/Z dataset (Telljohann et al., 2004), adapting English approaches that are trained on OntoNotes (Pradhan and Ramshaw, 2017). We intend to train and evaluate further on the DROC (Krug et al., 2018) and DraCor (Pagel and Reiter, 2020) datasets adapting our models to perform character based coreference resolution. In the context of event extraction, the focus on characters could benefit us by irrelevant events being discarded, on the other hand, the removal of non-character related events relevant to the plot (e.g.: an earthquake) could be detrimental.

Table 1 shows our best results for each model on the validation set (with which early stopping is performed). All models were tested in their base variant. We use the training, validation, and test splits suggested by Roesiger and Kuhn (2016). Multilingual BERT (Devlin et al., 2019) performs about on par with the two older German models but is outperformed by the more recently released Electra model[2].

On the test set, our approach also performs well, reaching an F1 score of 75.44 using the evaluation script by Pradhan et al. (2014). Existing German results on the same data, using the same prediction setup (i.e. without using gold mentions), reach a maximum F1 score of 48.54 (Roesiger and Kuhn,

2016). Our preliminary results show that the existing approach by Xu and Choi (2020) adapts well to German data, out-performing previous rule-based systems. We attribute this clear improvement over the current state of the art mostly to the improvements in word embeddings; previous approaches on German data have not made use of transformer-based models. Comparisons with English provide limited insight due to the difference in datasets.

In our context tuning coreference systems for precision could be an option, but it remains to be seen how this would affect overall performance.

4.4 Narrative Schemas

As mentioned in Section 1, as a first step a schema segmentation needs to be performed. From surface-level features (like paragraphs) to content-based ones (like perplexity of event sequence models), we will openly explore different approaches. The evaluation of segmentations will pose a challenge, due to the lack of evaluation data; we will start with manual evaluation, potentially extending it to metric-based evaluation later on. There is also the issue of unclear definitions of schema boundaries, it is not clear, for example, if a *social gathering* schema should contain events for transportation to said social gathering (recall the example in Figure 1).

When considering the document from the perspective of an entity e, we get a sequence of events $E_0^{\{e,...\}}$ through $E_n^{\{e,...\}}$ where each ellipsis in the superscript may represent any number of additional entities involved with the event. Splitting event chains from each entities' perspective (based on, for example, a sequence model's perplexity) could be a suitable first step in creating schemas, resulting in a set of event chains for each entity. The second step would then unify all event chains sharing common events into schemas. Taking a more global approach involving all events in sequence, in conjunction with the entities related to them will also be considered.

After segmentation, each individual chain will be processed into a single fixed-size vector representation. We intend to evaluate the naïve approach of averaging event representations. Sequence models, such as LSTMs (Hochreiter and Schmidhuber, 1997), will also be evaluated, training them on the narrative cloze task we hope to use their state vectors as representations for schemas. Such schema vectors would, ideally, be close, in vector space,

---

[1]Result from Roesiger and Kuhn (2016)
[2]https://huggingface.co/german-nlp-group/electra-base-german-uncased

to semantically similar schemas. Due to the presumed length of event sequences, we will focus on recurrent models that allow for arbitrary input sizes.

5 Conclusion

We proposed segmenting chains of events to form multiple schemas in long documents, mentioning different approaches to the representation of events and to their segmentation. Further, we discussed the options for representing schemas to allow for their analysis and thereby the comparison of different documents. An open question for us is if the two-layer approach to schemas and events is sufficient, if needed a hierarchical approach involving levels of schemas will be considered.

As part of the event extraction process in this thesis, work on both semantic frame induction and coreference resolution for German language content will be advanced. The representation of events using continuous frame embeddings is a new approach in the domain of information extraction.

Specifics of sequence modeling and feature learning on events are vague, iterations on the proposed concepts are planned. The open question of how exactly schemas boundaries are to be defined still needs to be explored.

We intend to help enable the computational analysis of literary texts. Schema representations may be used for finding previously hard to find similarities in different documents, whereas event features can be used to identify events that are important to the narrative. Statistical and machine-learning-based approaches to event modeling will advance the understanding of events in a domain that yet received relatively little attention.

Acknowledgements

References

Heinz Ludwig Arnold. 2009. *Kindlers Literatur-Lexikon (KLL)*. J.B. Metzler, Stuttgart/Weimar.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 969–974, Genoa, Italy. European Language Resources Association (ELRA).

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised Learning of Narrative Event Chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio, USA. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - ACL-IJCNLP '09*, volume 2, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

Dipanjan Das, Desai Chen, André FT Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational linguistics*, 40(1):9–56.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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 *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 837–840, Lisbon, Portugal. European Language Resources Association (ELRA).

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. Language, speech, and communication. MIT Press, Cambridge, MA.

Robert Gaizauskas and Yorick Wilks. 1998. Information extraction: beyond document retrieval. *Journal of Documentation*, 54(1):70–105.

Daniel Gildea and Daniel Jurafsky. 2000. Automatic labeling of semantic roles. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 512–520, Hong Kong. Association for Computational Linguistics.

Evelyn Gius, Svenja Guhr, and Benedikt Adelmann. 2020. d-Prose 1870-1920. Type: dataset.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with Aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A Knowledge-Enhanced Pre-training Model for Commonsense Story Generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska de Jong, and Emiel Caron. 2016. A survey of event extraction methods from text for decision support systems. *Decision Support Systems*, 85(C):12–22.

Richard Johansson and Pierre Nugues. 2007. Using WordNet to extend FrameNet coverage. In *Proceedings of the workshop on building frame-semantic resources for scandinavian and baltic languages, at NODALIDA*, pages 27–30, Tartu, Estonia.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77. Publisher: MIT Press.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for Coreference Resolution: Baselines and Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong. Association for Computational Linguistics.

Markus Krug. 2020. *Techniques for the Automatic Extraction of Character Networks in German Historic Novels*. Ph.D. thesis, Universität Würzburg.

Markus Krug, Frank Puppe, Fotis Jannidis, Luisa Macharowsky, Isabella Reger, and Lukas Weimar. 2015. Rule-based coreference resolution in German historic novels. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 98–104, Denver, Colorado, USA. Association for Computational Linguistics.

Markus Krug, Lukas Weimer, Isabella Reger, Luisa Macharowsky, Stephan Feldhaus, Frank Puppe, and Fotis Jannidis. 2018. Description of a corpus of character references in German novels-DROC [Deutsches ROman Corpus]. *DARIAH-DE Working Papers*, 27.

Seyed Abolghasem Mirroshandel, Gholamreza Ghassem-Sani, and Mahdy Khayyamian. 2009. Using tree kernels for classifying temporal relations between events. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, pages 355–364, Hong Kong. City University of Hong Kong.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California, USA. Association for Computational Linguistics.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. A Survey on Open Information Extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3866–3878, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Janis Pagel and Nils Reiter. 2020. GerDraCor-Coref: A Coreference Corpus for Dramatic Texts in German. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 55–64, Marseille, France. European Language Resources Association.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 Task-17: English Lexical Sample, SRL and All Words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland, USA. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Sameer Pradhan and Lance Ramshaw. 2017. OntoNotes: Large Scale Multi-Layer, Multi-Lingual, Distributed Annotation. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 521–554. Springer Netherlands, Dordrecht.

Behrang QasemiZadeh, Miriam R. L. Petruck, Regina Stodden, Laura Kallmeyer, and Marie Candito. 2019. SemEval-2019 task 2: Unsupervised lexical frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 16–30, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ina Roesiger and Jonas Kuhn. 2016. IMS HotCoref DE: A Data-driven Co-reference Resolver for German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 155–160, Portorož, Slovenia. European Language Resources Association (ELRA).

Wolf Schmid. 2014. *Elemente der Narratologie*, 3., erweiterte und überarbeitete auflage edition. De-Gruyter-Studium. de Gruyter, Berlin/Boston.

Jennifer Sikos and Sebastian Padó. 2019. Frame identification as categorization: Exemplars vs prototypes in embeddingland. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 295–306, Gothenburg, Sweden. Association for Computational Linguistics.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 2004. The Tüba-D/Z Treebank: Annotating German with a Context-Free Backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 2229–2232, Lisbon, Portugal. European Language Resources Association (ELRA).

Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 2020. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8519–8526, Online. Association for Computational Linguistics.

Patrick Xia, João Sedoc, and Benjamin Van Durme. 2020. Incremental neural coreference resolution in constant memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8617–8624, Online. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533, Online. Association for Computational Linguistics.

Zheng Xin Yong and Tiago Timponi Torrent. 2020. Semi-supervised deep embedded clustering with anomaly detection for semantic frame induction. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3509–3519, Marseille, France. European Language Resources Association.