Narrative Cloze as a Training Objective: Towards Modeling Stories Using Narrative Chain Embeddings

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

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

Abstract

We present a novel approach to modeling narratives using narrative chain embeddings. A new dataset of narrative chains extracted from German news texts is presented. With neural methods, we produce models for both German and English that achieve state-of-the-art performance on the Multiple Choice Narrative Cloze task. Subsequently, we perform an extrinsic evaluation of the embeddings our models produce and show that they perform rather poorly in identifying narratively similar texts. We explore some of the reasons for this underperformance and discuss the upsides of our approach. We provide an outlook on alternative ways to model narratives, as well as techniques for evaluating such models.

1 Introduction

The narrative cloze task was originally introduced by Chambers and Jurafsky (2008) and is the task of, given a sequence of narrative triples, predicting a masked triple. Such triples are made up of subject, verb, and object, and the triples in one chain share a common participant, referred to as the protagonist. Their subsequent work (Chambers and Jurafsky, 2009) improved upon the results from the original paper and formulated the task slightly differently, expanding it to schemas with multiple participants. Granroth-Wilding and Clark (2016) extract additional information and introduce evaluation metrics. An excerpt from one of their automatically extracted chains goes as follows: (A, plead, [with, B]), (_, heartbroken, A), (A, die, _), where A and B are entities, and each triple represents a verb with its arguments.

One of the early motivations for the narrative cloze task was modeling narrative contexts and inferring narrative schemas (Chambers and Jurafsky, 2009). We aim to adapt the semantic modeling performed as part of the task to identify documents that share similar narrative schemas rather than explicitly inferring such schemas. That is to say: analogously to masked language modeling, we use narrative cloze as a training objective to train narrative understanding, rather than language understanding. The motivation being that abstract story similarities may be found, eventually enabling computational comparisons of stories rather than texts. Such an approach could, for example, be useful in digital humanities with researchers already experimenting with word embeddings to identify and compare adaptations of the same story (Glass, 2022). Our approach to modeling narratives constitutes a continuous and embedding-based approach to schemas like Propp's model of Russian folklore (Propp, 1968). The chain-based approach has the upside of allowing for abstracting over information that is not relevant to the actual narrative, but that will be captured by more recent semantic embedding methods like SentenceBERT (Reimers and Gurevych, 2019). The method's potential downside, however, as discussed by Wilner et al. (2021) is that too much contextual information is lost, making the task of predicting triples ambiguous or impossible. Through the use of contextual embeddings and an optional additional re-contextualization process, they improve on existing narrative cloze results by using additional information. Our ultimate goal of this work is to enable embedding-based computational narrative similarity comparisons of texts, a task we see as closely related but not identical to the popular narrative generation field (see e.g. Gervás, 2021).

The three key contributions of this work are **(1)** a dataset of German narrative chains and **(2)** the application of narrative embeddings to a down-stream task in the form of replicating human narrative similarity judgments, as well as **(3)** state-of-the-art models on English and German for narrative chains without external information from contextual embeddings.

Proceedings of the The 5th Workshop on Narrative Understanding, pages 118–127
July 14, 2023 ©2023 Association for Computational Linguistics

2 Background

To evaluate the capability of our embeddings in recognizing similar narratives, we rely on comparisons to human annotations. Conceptual work on text similarity (Bär et al., 2011) pointed out that text similarity is not inherently well defined by showing that, without further instructions, some annotators focus strictly on content, whereas others additionally take the text's structure into account. Accordingly, our task calls for a dataset that explicitly annotates narrative schema similarity. Chen et al. (2022a) introduced such a dataset in the form of a multilingual news similarity dataset containing the similarity of news article pairs along seven dimensions. According to their annotation code book (Chen et al., 2022b), dimensions are to be rated independently of each other, with the *narrative* dimensions focusing on similarity in narrative schemas as defined by Chambers and Jurafsky (2009); the dataset thus contains human ratings of schema similarity.

Since its inception, the narrative cloze task has seen work in different directions. Chambers (2017) has criticized newer approaches to the task as deviating from its original formulation, focusing on extracted events in text order rather than manually annotated ones; they emphasize that the automated approach is much more aligned with the capabilities of language models. Wilner et al. (2021) approach the narrative cloze task but reformulate it to use contextual embeddings instead of verb lemmas. While this approach yields much higher accuracies and can help disambiguate events, we feel that in the light of modeling narrative disjointly from the surface form, such contextual embeddings would potentially hamper the model's performance in any downstream application.

In the narrative cloze task, the model is asked to predict a masked triple describing an event. In practice, this is a four-tuple of the subject, verb, indirect object, and object in more recent implementations like the one by (Granroth-Wilding and Clark, 2016). Evaluation has, as suggested by Granroth-Wilding and Clark (2016), in the recent past been performed in a MCNC (multiple choice narrative cloze) setup where the model is asked to pick the most fitting triple for a corresponding masked triple in a chain given exactly 5 options that are randomly sampled from the entire corpus. This evaluation setup was introduced to enable more interpretable results and pays tribute to the fact that, in most cases, the exact triple is ambiguous not just by virtue of synonymous verb lemmas but also due to contextual ambiguity.

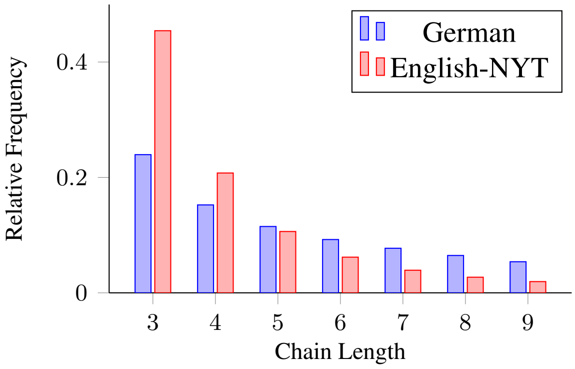

Figure 1: Relative distribution of chain lengths our German dataset compared to the English NYT dataset

The work by Granroth-Wilding and Clark (2016) discusses multiple models, with the best score being achieved by a model that calculates the compatibility of a given candidate triple by averaging across its compatibility (as scored by a neural model) with all other elements of the chain.

3 Datasets

We use the Gigaword dataset with the preprocessing pipeline presented by Granroth-Wilding and Clark (2016). In addition, we build a German dataset based on scraped German news data. The data is extracted using a German coreference resolution system by Schröder et al. (2021) and dependency parsing from SpaCy (Honnibal et al., 2020). We produce a dataset of around 1.8×10^6 German narrative chains; we filter out any chains shorter than three at dataset creation time. Compared to the approximately 5.7×10^6 chains with a length of at least three in the Gigaword-derived dataset, this is a relatively small collection but still allows us to explore the adaption to a different language.

While we rely on the intrinsic MCNC evaluation for comparison to existing work, for assessing the use for narrative modeling, we need a downstream evaluation, and only limited data is available for this purpose. The multilingual news-similarity SemEval dataset (Chen et al., 2022a) is, at first sight, a great fit; the pairs of articles making up the dataset are each annotated with regard to their similarity along seven specific dimensions, with each being dimension scored on a scale of 1–4. The dataset's *narrative* dimension is, however, highly (ρ=.88) correlated with its *overall* dimensions, meaning that when articles are narratively similar, they are likely to also be similar in a general sense. It may seem that, due to this alignment in similarity, no differentiation needs to be made for modeling the two dimensions, but we believe this difference is crucial in identifying texts that deal with the same narrative in different circumstances. This difference may also be interesting for other domains, especially narrative literary texts, where the correlation may, in practice, not be as high. In these texts, two scenes telling a similar narrative may not share any concrete entities; for example, the circumstances of two arguments between multiple characters may be entirely different with different surroundings and differently named characters, yet share some conceptual similarity. As the *overall* dimension can be modeled well using existing text similarity models, however, it seems unlikely that our approach based on narrative chains will be able to outperform existing models for the news domain. Still, we employ the dataset as a testbed for extrinsic evaluation for the narrative cloze task.

We make all our extracted chains, the ones from the NYT dataset, our German dataset, and the SemEval dataset, available for download to enable further research. [1]

4 Experimental Setup

To enable some comparison with prior work, we replicate the testing setup by Granroth-Wilding and Clark (2016) wherever possible. In this section, we discuss the specifics of the task and provide an embedding baseline for our downstream evaluation.

4.1 Task Details

Various evaluation details for the MCNC are not clearly defined; subsequently, we discuss the parameters we chose as well as their impact on the evaluation.

Minimum Chain Length: Chambers and Jurafsky (2008) only consider chains of a length of at least five triples. While Granroth-Wilding and Clark (2016) do not explicitly discuss this parameter but seem to also apply a limit, the exact value is not known to us; in the implementation, a default value of 9 is present. A minimum length limit seems reasonable as (a) predicting lemmas in chains of length one is largely up to chance, and (b) an actual story is likely told with multiple events. In line with (a), we found the choice for this evaluation parameter to have a fairly large impact on our results; for example, using the minimum chain lengths 9, 5, and 3 resulted in the accuracy dropping from 50.21 to 49.08 and 48.48 respectively for a variant of our static embedding model on the dev set. Choosing a specific value is, to some degree, an arbitrary decision; for comparability, we adopt the choice of a minimum chain length of 9 in our experiments.

Minimum Lemma Count: With this parameter, verb lemmas below a certain absolute count are removed from the training and evaluation data. Due to the long-tail nature of verb lemma count distributions, many verbs occur very infrequently in the input. In preliminary experiments, we found this to have some impact on the results; it is not clear which threshold was chosen in previous work. We do not employ this filtering step and instead use all verb lemmas that occur in the dataset.

Maximum Lemma Count: In previous work, "stop events" have been used to refer to the process of excluding verbs that occur too often. Rather than picking a specific threshold in terms of count, Granroth-Wilding and Clark (2016) used the top ten most frequent verbs. We found this filtering criterion helpful for model convergence (otherwise, the very frequent lemmas would dominate others). While "see" or "go" are not stop words in the traditional sense (i.e., they do carry semantic information in a text), in the context of our chains, in the news domain, they could conceivably occur in any chain at any point and do not bear any information content.

Chain vs. Schema-based: Evaluation can either be performed on the basis of entire narrative schemas, i.e., multiple chains that share common participants or on the individual chain. In this work, we operate on individual chains making the multiple choice task, at least in theory, harder than in full-schema scenarios.

Mention Surface Forms: Including the surface form of entities means including the concrete form of each entity mention in the triple. Consider the short chain (A, gives, B) (B, write, C) and compare it with a version including surface forms (A: source, give, B: reporter) (B: reporter, write, C: article). Here predicting the second verb is difficult with no entity surface forms given, but once entity information is present, the task becomes manageable. In an open prediction task without multiple choices, the solution only becomes relatively un-

[1] https://ltdata1.informatik.uni-hamburg.de/narrative-chains/

ambiguous when the surface form “article” is also given.

Candidate Triples: Another important parameter is the makeup of the triples the model is asked to choose from in the MCNC evaluation. In terms of candidate triple selection, Granroth-Wilding and Clark (2016) randomly sample from all triples in the dataset, as setup which we follow. The second aspect is whether the whole triple is presented as a candidate solution, which is largely the case in prior work, although Wilner et al. (2021) also consider a verb only variant. It is clear that with actual full text for the events (i.e., the mention’s surface forms), the prediction is trivial in many cases, as entity names are usually unique within the five presented choices. For this reason, we only mask the verb in our experiments (except for when explicitly stated in the case of the T5 model, see below), sampling four random verb lemmas from the dataset as the distractors in the MCNC task.

4.2 Downstream Evaluation and Baseline

We perform the extrinsic evaluation on narrative similarity (using the dataset by Chen et al., 2022a) by means of embedding similarity. To align with their evaluation and following a substantial number of submitted systems in their shared task, we embed each document independently and compute the cosine similarities.

			Dimension	
Model	Dataset	Overall	Narrative	Entity
All Verbs	EN	49.40	**50.02**	50.58
All Words	EN	43.12	43.37	44.10
Chain Verbs	EN	19.99	19.21	14.03
Chain Mean	EN	12.65	11.09	6.63
Transformer[2]	EN	81.78	**78.16**	83.76
Chain Verbs	DE	44.81	**48.49**	41.07
Chain Mean	DE	24.56	19.12	17.91

Table 1: Correlation of cosine distance of fastText embeddings with the dimensions *overall* and *narrative* on the English evaluation split of Chen et al. (2022a), with a sentence transformer model provided as a comparison.

As a weak baseline for comparing narratives, we introduce a word embedding-based comparison. For simplicity, we only consider those pairs where both articles are written in our model’s language (either English or German). On the English and German sections of the news similarity evaluation data, we compute fastText (Bojanowski et al., 2017) embeddings of all words, all verbs, and then of all the verbs included in the narrative chains. The best results were achieved using a word-level best-match approach, following BertScore’s (Zhang et al., 2020) token similarity matching. For comparison, we also provide a method where this matching is done on the mean of the verb embeddings of individual chains and, therefore, a chain best match approach. Table 1 shows that these approaches lack far behind a sentence encoder baseline and that while a focus on verbs helps, especially concerning the *narrative* dimension, the limitation of only including the verbs that are part of narrative chains as extracted by Granroth-Wilding and Clark (2016) pipeline severely impacts the results. We can observe that, for the German evaluation split, the results are generally much better than for the English data. We attribute this to the improved extraction pipeline. Note that we discard all pairs where either document has no extracted chains; unlike in the German training dataset, even chains of a length below three are retained. Taking only the verb embeddings clearly outperforms the variant that considers all words; we do not even see a clear effect concerning the narrative dimension being represented better by this setup. Given these initial results, it seems possible that the “all verbs” embedding baseline will not be outperformed. Nevertheless, it remains interesting to see if the narrative cloze task can prioritize the narrative dimension over others.

Embeddings	MCNC
FastText-German	**31.23**
Muse-German	25.04
BPEmb	30.19

Table 2: Comparing embedding sources on the German dev set. No mention surface forms are used.

5 Model Setup and Architecture

We present a neural model that, using static word embeddings as input features, performs state-of-the-art narrative cloze prediction.[3] To provide an additional point of comparison, we build a baseline based on modern techniques, specifically the T5 (Raffel et al., 2020) architecture and training setup.

[2] We use all-mpnet-base-v2 from Reimers and Gurevych (2019).

[3] Implementation: https://github.com/uhh-lt/narrative-chain-embeddings

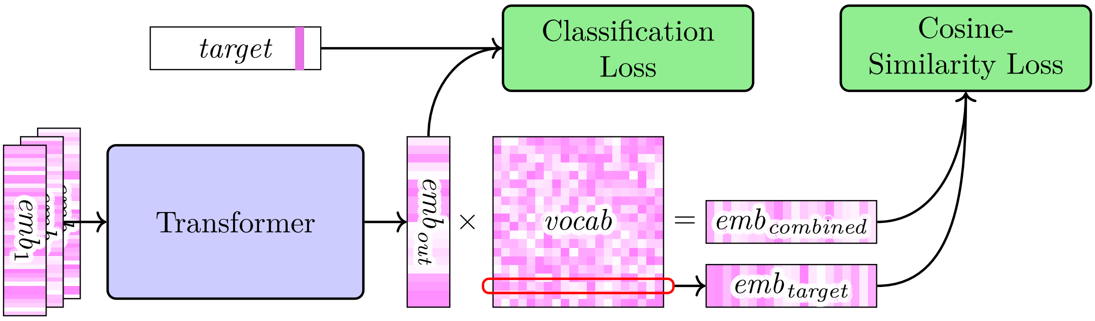

Figure 2: In our model architecture, we improve training using a linear combination of embeddings in the output vocabulary.

5.1 Static Embedding Approach

We model narrative sequences using a fixed-sized context of surrounding triples. Our model is a transformer that makes use of static embeddings of individual words as the input (cf. Fig 2), unlike Granroth-Wilding and Clark (2016), we do not train static embeddings based just on verbs but instead rely on existing embeddings trained on entire texts. We take a twofold approach to entity representation, allowing both word embeddings of entity surface forms as well as identity one-hot-encodings that remain consistent inside of a specific chain. These entity representations are concatenated with the verb lemma's embedding to form our model's input embedding for each triple. The training objective, inspired by masked-language-modeling, has the model predict one verb lemma at a time. Due to the long tail distribution of verb lemmas, we need a fairly large but manageable output vocabulary of $\approx$7500 words for the English Gigaword-based dataset.

Our model approaches the task as a classification task at inference time, in that the output is a probability distribution across the vocabulary. To improve convergence, we train on a cosine distance objective; the loss function is a cosine-similarity-based embedding loss, comparing the output-distribution-weighted average of the classes' word embeddings with the gold class's corresponding word embedding. The more straightforward approach of using a cross-entropy classification loss did not produce adequate results. During training, we do not update any parameters in the system creating the embeddings. We expect that the embedding loss allows us to learn better from ambiguous training examples, as the embeddings of semantically similar verbs will also have a smaller cosine distance. For extrinsic evaluation, we use the emb_{out} embedding, the output state of the transformer.

In terms of embedding sources, Table 2 shows a minimal difference between BPEmb (Heinzerling and Strube, 2018) and FastText for German, making BPEmb an interesting choice and possibly enabling cross-lingual knowledge transfer.

For all presented training runs on the static embedding approach, we use the same set of manually optimized hyperparameters: a dropout chance of 0.2, a learning rate of 1×10^{-3}, and the one cycle learning rate scheduler (Smith and Topin, 2019). The scheduler increases the learning rate for the first 30 epochs, slowly decreasing it afterwards. In practice, early stopping finished most runs shortly before or after reaching the maximum learning rate.

5.2 Langauge Model Approach

For comparison, we employ a state-of-the-art language model in the form of T5, converting chains into textual representations of the form `(subj, verb lemma, iobj, obj)`, where subject, object, and indirect object each come with a unique identifier and the mention's surface form. We use the tiny variant of T5 with randomly initialized weights with a custom tokenizer trained on our dataset. Our implementation is based on an existing training script, meaning the masking is not limited to verbs but instead to random tokens in the input.

For the MCNC task, we align the inference with T5's denoising training objective by masking a single event and comparing the likelihood of all multiple-choice options as generated outputs. Embeddings are created by using the last encoder state of the T5 model.

Setup	MCNC
Full Model	**52.00**
+ classification loss	48.29
+ classification loss - embedding loss	25.04
- mention surface forms	50.21
- mention surface forms - FastText + BPEmb	49.78

Table 3: Ablation study for our model on the English dev set with our static embedding approach, + and - indicate added and removed model options, respectively.

Dataset	Entity String	Model	MCNC-Accuracy
Ours (German)	✗	Ours	30.66
Gigaword-Verb	✓	Ours	**50.89**
	✗	Ours	49.06
	✗	T5-based	28.01
Gigaword-Triple	✓	T5-based	**92.33**
	✓	G&C (2016)	49.57
Gigaword-Context	✗[4]	W,W&G (2021)	92.22

Table 4: MCNC results on the Gigaword NYT and our own dataset show that our models outperform previous approaches in the same setup.

Model	Dimension		
	Overall	Narrative	Entity
Ours (no surface forms)	11.06	16.68	10.82
+ shuffle	11.33	**17.18**	10.65
- entities	8.76	11.83	7.26
Ours German	25.78	23.64	21.64
+ shuffle	25.71	**23.94**	21.55
- entities	26.74	23.66	21.73
English T5 model	13.17	**9.95**	10.13
+ entity surface forms	5.69	2.53	6.05

Table 5: The extrinsic evaluation on the news similarity dataset is evaluated using Pearson correlation of embedding distances with human judgments.

6 Results

In Table 3, we report the impact of different parameters on our model. In terms of embeddings, FastText slightly outperforms BPEmb by .43 percentage points but does not provide any multilingual capabilities. Additionally, the impact of mentions' surface forms is only 1.79 percentage points, making it potentially viable to exclude them, thereby increasing the model's focus on the narrative over the mentioned entities. For the choice of loss functions, it is clear that the embedding loss performs much better than the classification-based loss on its own by a large margin of 26.96 percentage points; even the combination of both performs appreciably worse than the embedding loss on its own.

We did not find success with reusing weights, from our BPEmb setup, from one language in the other but did not experiment with multi-task learning to handle both languages at once.

Table 4 shows that our model outperforms previous approaches in the MCNC setting with a minimum chain length of nine, outperforming approaches in the same setup by more than 1.8 percentage points. Further, it shows that the inclusion of entity surface forms enables the T5 model to perform incredibly well at over 92% accuracy, making it ostensibly outperform the best models by Wilner et al. (2021), which uses contextual representations. Their evaluation setup may, however, differ in terms of minimum chain length, making this comparison an unclear one. It is to be noted that the Gigword-Triple models are asked to predict the entire triple of arguments rather than just the verb lemma, as is the case for the other models. As supported by the much worse performance of the T5 model without access to entity strings (a drop by over 60 percentage points), we strongly suspect that the T5 model is only looking for compatible mentions and will often only find one such option in the five choices presented. We manually confirmed that this strategy works in the majority of cases. The performance compared to that of the Granroth-Wilding and Clark (2016) model can be explained by the fact that this model only compares pairs of triples, averaging across their coherence scores, and can thus not look for mention compatibility globally in the entire chain. Overall, removing entity surface forms leads the T5 model to underperform drastically, whereas our static-embedding-based model only suffers a minor performance penalty. As previously discussed, we suspect this setup may lead to more meaningful narrative modeling.

As a downstream evaluation of our embeddings, in Table 5, we use them to predict the narrative similarity as annotated by humans in the multilingual news similarity dataset (Chen et al., 2022a). Our results clearly show that narrative chains fail to be a good model of narrative, with our results on static embeddings indicating that the loss of context is, at

[4] While the model does not explicitly use the mention's surface forms, they are captured by the verb's contextual embedding.

least in part, at fault. Table 5 further supports our explanation of T5's overperformance; rather than focusing on semantic aspects of the chain, T5 appears to focus on matching mention surface forms, which is reflected in its very low performance on the extrinsic evaluation.

After qualitative analysis, we suspected that our model might only be a topic model of sorts that considers the domain of verbs rather than any sequential nature of them. This is supported by the fact that it is overall still comparable in MCNC performance with the coherence based Granroth-Wilding and Clark (2016) model. Further news articles often do not tell happenings in their chronological order while our extraction pipelines rely on text order, meaning that the order does not necessarily follow logical sequences of actions. We test this hypothesis of no sequential understanding in Table 5 by shuffling the triple sequence. We find that both models perform slightly better with shuffling on this specific data (although only by a margin of up to 0.5 percentage points), proving that there is, in fact, no reliance on ordering information. Interestingly, removing entities (meaning identity information in the form of one-hot encoding rather than surface forms in case) has a much larger impact of ≈5 percentage points on the results for the English dataset. This is in line with our findings in manual prediction experiments on the MCNC task, where we found a good strategy to be the compatibility of actions of a given entity (e.g., someone who "raises" may also "announce" or "purchase" but probably will not "live"). The effect of entities having a large effect on the results is, however, not seen in the German data, indicating that it may take a different approach to narrative modeling. Overall the German model exhibits better performance, which may be attributed to the different extraction pipelines, which already produced better results in Table 1; in fact, the German model is the only one that outperforms one of its baselines, the "Chain Mean" variant by a margin of ≈4 percentage points on the narrative dimension. This is surprising, given that it performed much worse than the other variants on the MCNC, casting doubt on the usefulness of narrative cloze evaluation, at least in this specific setup.

6.1 Silhouette Scores

To further inspect the model, we analyze the produced embeddings in terms of their clustering. Specifically, we make use of the silhouette score (Rousseeuw, 1987), a cluster evaluation metric, to assess how well-separated individual verb lemmas are. The embeddings are created by masking an individual lemma and taking the corresponding predicted output representation. The silhouette score can take values from -1 to 1, where each of the extremes means the data points are perfectly mixed and perfectly separated. To be clear, we do not expect a perfect performance from either method here, as polysemous verbs mean that the same lemma should not always receive the same embedding while (due to synonyms) different lemmas may take the same embedding form; a comparison across models may, however, provide additional insights.

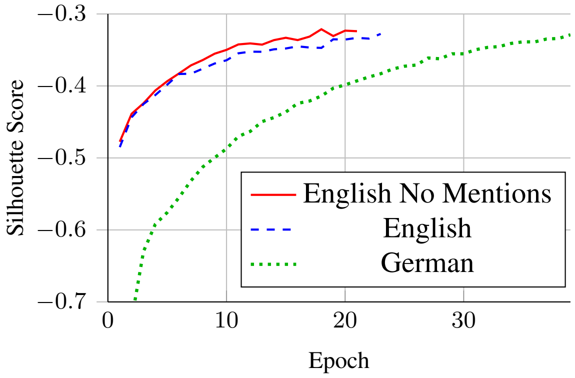

Figure 3: Silhouette scores of three models keep improving throughout training, indicating that verb lemmas are increasingly separated throughout the training process.

Figure 3 illustrates that, in all runs, the silhouette scores steadily improve. For the German dataset, it is expected that convergence takes more epochs due to the smaller training set, but it is surprising that the silhouette score ends up at -0.33, equivalent to both English runs at -0.33 and -0.32, despite the much worse performance of the models on the MCNC task. This result further supports the idea that the narrative cloze task, in its current form, may not be a perfect approximation of narrative modeling capabilities.

6.2 Qualitative Exploration

For insights into the model's performance, we manually assess its output. First, we ask the English model (without mention surface forms) to predict the lemma in a triple of two participants. The model outputs the following lemmas: "join", "win", and "support". Interestingly, this is not in line with the most common lemmas ("think", "play", and "call", after filtering stop-lemmas), which may be explained by short chains having different content than longer ones.

The chain (A, kill, B), (C, catch, A), (D, _, A), where the underscore denotes a masked lemma, results in the following top three lemmas predicted in descending order of probability: "hit", "find", "face". If we add the information that we are in a judicial context by adding (D, sentence, A) to the end of the chain, we get the following list of lemmas instead: "shoot", "kill", "catch". While these lemmas are more compatible with the domain, it seems unlikely that the same entity sentencing the subject would also shoot or kill them, indicating that the identity information of entities does not have a large effect.

We test if the ordering can, in extreme cases, affect the outcome using the following chain: (A, hug, B), (A, insult, B), (A, _, B). In this chain, changing the order of "hug" and "insult" leads to the lemmas "kiss" and "hit" changing their order in terms of model score, with "hit" receiving the higher score when "insult" comes directly before it. This reversal indicates that some ordering information is present in the model even though it is not conducive to narrative embeddings (as evidenced by the results in Table 5). We observe the same behavior in the German model using a translation of the above chain.

The examples illustrate the natural ambiguity created by removing most contextual information, an effect that likely places an upper limit on MCNC performance. The fact that ordering information is used to check the compatibility is a promising sign that some narrative understanding may be present in our model that goes beyond the best-performing approach by Granroth-Wilding and Clark (2016), which does not take order into account.

7 Conclusion

In this work, we presented models with state-of-the-art MCNC performance in two different setups and on German and English datasets. We produced vector embeddings as narrative representations and performed extrinsic evaluations of our narrative cloze models using the comparison to human narrative similarity ratings. In a qualitative review of our model outputs, we illustrated that the model captures sequential information. The performance of our embeddings indicates that narrative-cloze may not be a perfect fit for narrative similarity modeling; on the other hand, we were able to, in some scenarios, produce embeddings that model narrative similarity better than overall similarity, placing emphasis on the desired aspects of a text. In almost all cases, our models were also able to place less emphasis on entities than plain word embedding and especially sentence encoder models did. Overall, it can be concluded that limiting the model's access to information can help create embeddings that represent a specific aspect of the text.

It is also clear that in the current state, in almost all setups, our chain embeddings are outperformed even by static verb embeddings. We see two major roadblocks to applying this approach to the computational modeling of narratives. The first is the limited evaluation data: while the SemEval dataset by Chen et al. (2022a) is a step in the right direction, it fails to clearly demonstrate the need for narrative modeling as, in the news domain, dimensions are strongly correlated. A dataset on another domain is needed; this is something we seek to address in upcoming work.

The second is the actual quality of predictions. In preliminary annotation experiments, we were unable to perform on par with the predictions system. While further analysis is required, we suspect that this is attributable to the fact that the chains provide too little information.

8 Future Work

As we see the limited information as a crucial shortcoming of narrative chains, we will conduct further research in the direction of Wilner et al. (2021), using contextual embeddings and trying to explicitly remove information on the actors (e.g., by renaming them). In our opinion, the approach of narrative cloze in its original form is no longer a promising approach for building semantic representations of narratives. Avenues to improving the performance on the narrative cloze task still exist and go beyond improving the extraction process or the representation of individual events. An example of this may be exploiting the knowledge of pre-trained large language models, which we did not find success in preliminary experiments.

If the semantic modeling by means of extracted narrative chains was to be successful in the future, we suspect that a much-improved event representation would be needed. It may, however, be more promising to pursue alternative ways of modeling narratives, perhaps through the use of supervised narrative similarity data. Any supervised training on the text level will, however, need to deal with the effect that other similarity markers, such as common entity names, already are a strong indicator of narrative similarity. Such markers are not present during inference on unrelated texts sharing similar narratives.

Acknowledgements

This work was supported by the DFG through the project “Evaluating Events in Narrative Theory (EvENT)” (grant BI 1544/11-1) as part of the priority program “Computational Literary Studies (CLS)” (SPP 2207).

References

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2011. A Reflective View on Text Similarity. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 515–520, Hissar, Bulgaria. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Nathanael Chambers. 2017. Behind the Scenes of an Evolving Event Cloze Test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

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.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022a. SemEval-2022 Task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, Washington, USA. Association for Computational Linguistics.

Xi Chen, Ali Zeynali, Chico Q. Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw A. Grabowicz, Scott A. Hale, David Jurgens, and Mattia Samory. 2022b. SemEval-2022 Task 8: Multilingual news article similarity: Codebook for text similarity annotations. URL: `https://zenodo.org/record/6507872`.

Pablo Gervás. 2021. Computational Models of Narrative Creativity. In Penousal Machado, Juan Romero, and Gary Greenfield, editors, *Artificial Intelligence and the Arts: Computational Creativity, Artistic Behavior, and Tools for Creatives*, pages 209–255. Springer International Publishing, Cham.

Grant Glass. 2022. An Adaptive Methodology: Machine Learning and Literary Adaptation. In *Digital Humanities. 2022 Combined Abstracts*, pages 210–212, Tokyo, Japan.

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? Event prediction using a compositional neural network model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2727–2733, Phoenix, Arizona, USA.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2989–2993, Miyazaki, Japan. European Language Resources Association (ELRA).

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. Software Release, URL: `https://zenodo.org/record/7970450`.

Vladimir Iakovlevich Propp. 1968. *Morphology of the Folktale*, volume 9. University of Texas Press.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):140:5485–140:5551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Fynn Schröder, Hans Ole Hatzel, and Chris Biemann. 2021. Neural End-to-end Coreference Resolution for German in Different Domains. In *Proceedings of* *the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 170–181, Düsseldorf, Germany. KONVENS 2021 Organizers.

Leslie N. Smith and Nicholay Topin. 2019. Super-convergence: very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, Baltimore, Maryland, USA. International Society for Optics and Photonics, SPIE.

Sean Wilner, Daniel Woolridge, and Madeleine Glick. 2021. Narrative Embedding: Re-Contextualization Through Attention. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1405, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Eighth International Conference on Learning Representations*.