Tell me again! A Large-Scale Dataset of Multiple Summaries for the Same Story

Hans Ole Hatzel, Chris Biemann
Universität Hamburg,
{hans.ole.hatzel,chris.biemann}@uni-hamburg.de

Abstract

A wide body of research is concerned with the semantics of narratives, both in terms of understanding narratives and generating fictional narratives and stories. We provide a dataset of summaries to be used as a proxy for entire stories or for the analysis of the summaries themselves. Our dataset consists of a total of 96,831 individual summaries across 29,505 stories. We intend for the dataset to be used for training and evaluation of embedding representations for stories, specifically the stories' narratives. The summary data is harvested from five different language versions of Wikipedia. Our dataset comes with rich metadata, which we extract from Wikidata, enabling a wide range of applications that operate on story summaries in conjunction with metadata. To set baseline results, we run retrieval experiments on the dataset, exploring the capability of similarity models in retrieving summaries of the same story. For this retrieval, a crucial element is to not place too much emphasis on the named entities, as this can enable retrieval of other summaries for the same work without taking the narrative into account.

Keywords: story, sentence embeddings, summary, representation learning

1. Introduction

The area of representing the semantics of narratives has received considerable attention (Chambers and Jurafsky, 2008, 2009; Granroth-Wilding and Clark, 2016). Previous work has typically operated in the domain of news texts, rather than on fictional narratives. In this work, we seek to represent stories in the wider domain of long-form narrative texts. Semantic modeling of stories is often performed in the context of story generation (e.g. Rishes et al., 2013), where it typically operates on fictional narratives.

The semantics of specific narratives are hard to explicitly define and, while limited domain approaches exist (Propp, 1968) and attempts to automate their annotation have been made (Finlayson, 2012), a generalized approach based on explicit schemas does not seem feasible as narrative semantics represent an open class. A more flexible semantic schema approach based on interaction triples by Chambers and Jurafsky (2008) was, with limited success, recently applied to narrative similarity search (Hatzel and Biemann, 2023).

With this work, we want to pave the way for supervised representation learning of semantic narrative representations. In a simple setup, this could mean using contrastive learning to produce similar embeddings for similar stories. This approach is currently limited in the availability of training data, which is hard to create for this task as annotators can not find similar or identical stories across a large dataset. At the same time, sampling random pairs of stories will yield, in the vast majority of cases, stories that are not even marginally similar. Chen et al. (2022), who annotate news stories with regard to their similarity in terms of narrative schema, approach this by, among other techniques, relying on vocabulary-based Jaccard similarity in document pairs, thereby collecting pairs that often also share some semantic overlap. As a result, pairs annotated as similar typically also exhibit a considerable vocabulary overlap, meaning that the training data contains virtually no instances of stories with different entities and settings being considered similar in terms of the narrative. Quantitatively, this is expressed in high correlations between, for example, the narrative similarity and the similarity in the occurring entities. In effect, representation learning will thus fail to learn the actual similarity of stories, capturing such spurious correlations as that of entities with narrative instead.

Limited work addressing the semantics of long-form narrative texts is available, with many approaches focusing on more structured texts like screenplays (Gorinski and Lapata, 2018; Bhat et al., 2021). We propose that, due to the inherent complexities associated with long narrative texts, summaries can reasonably be used as a stand-in for long stories to advance research in this direction. Operating on summaries can enable downstream applications as human-written summaries of texts are often available. Further, it is to be expected that automated summarization of long-form texts will also improve over time. Summaries can further provide the benefit of being under permissive licenses even when the original work is not, allowing for freely available academic datasets.

We present a dataset[1] of almost 100,000 story summaries scraped from Wikipedia with metadata

[1] https://github.com/uhh-lt/tell-me-again

LREC-COLING 2024, pages 15732–15741

retrieved from Wikidata, a community-maintained knowledge graph. Our dataset includes summaries of books as well as movies and is not limited to fictional works, including such works as biographies. The dataset is further distinguished in that all entries come with Wikidata IDs, allowing downstream users of the dataset to incorporate a wide range of additional information. For example, metadata can, in many cases, enable the retrieval of the respective full text of summarized books, e.g. by purchasing the book with the respective ISBN. Applications for this metadata could also be found in the field of Digital Humanities: for example, in computationally analyzing summaries, either as a proxy for the actual text or for their own sake, associated with specific authors or genres.

2. Related Work

In terms of summary datasets, we are aware of two datasets closely related to this work. WikiPlots[2] contains over 100,000 plot summaries extracted from only the English Wikipedia without additional metadata (besides titles). Chaturvedi et al. (2018) released a dataset similar to ours, based on the idea of collecting summaries from movies and their remakes. This dataset was harvested based on a Wikipedia list of movie remakes. The authors provide a total of 266 clusters of remakes, where each cluster typically contains summaries of two different movies, with 31 clusters containing three or more movies and their summaries (see Figure 4 for a visual breakdown). While this dataset is a good fit for evaluation, it lacks the scale required for most supervised learning methodologies, as also evidenced by their decision to test on 80% of their data. They also introduce a story kernel designed to "quantify narrative similarity" by means of the similarity of both plot (as modeled by events and entities) and characters (i.e. persons in the story). We will use their dataset as a comparison in the experiments on our newly created dataset.

Wu et al. (2021) summarize long documents by training language models to first summarize small sections of the original text and then combine multiple such summaries into a new one, recursively building up to a single summary for an entire text. In the process, they create human-curated summaries for books but only release small samples of their data. Kryściński et al. (2022) provide a dataset of 217 books with a total of 405 summaries that are each aligned to the source text on a paragraph level. The dataset contains on average around 1.8 summaries per book and was created based on the idea of aligning summaries with their original texts.

The ROCStories dataset by Mostafazadeh et al. (2016) provides an evaluation testbed that asks models to predict which of two endings matches a story. This dataset was designed to measure common sense reasoning and makes use of short 5-sentence stories written specifically for the dataset.

Lastly, we want to mention the WikipediaSummary-Dataset (Scheepers, 2017), which is similar to ours only in name. It uses the first paragraph of Wikipedia pages in conjunction with the full article as data for text summarization.

3. Dataset Creation

To ensure well-structured metadata, we start our dataset creation process on the basis of Wikidata. We query Wikidata for all literary works and movies that have links to Wikipedia pages in German and English. From the linked Wikipedia pages, we extract, based on their headings, the section that contains a content summary.

We retrieve a set of summaries of the same work from multiple language versions of Wikipedia (considering the English, German, Italian, French and Spanish Wikipedia language versions). As we intend this dataset to primarily serve as a training dataset for narrative similarity, we suspect that, while recent multilingual models can handle such data, that language differences may complicate training (e.g. no two documents of the same language in the training data would ever describe the same story). Additionally, we use data augmentation techniques (see Section 5.1) which are much more labor-intensive to apply in a multilingual setting. For these reasons, we also provide machine-translated English versions of all summaries. We use a pre-trained machine translation model (NLLB-Team et al., 2022), specifically the 3.3 billion parameter variant in conjunction with a ready-to-use application (García-Ferrero et al., 2022). For translation from the four source languages to English, NLLB-3.3B achieves competitive results with the much larger MoE (mixture of experts) model and is one of the strongest freely available translation models. In our qualitative evaluation, although we did spot occasional errors, the translation quality was adequate. We deliver the dataset with all the original texts, enabling future translation with a stronger system if desired.

Figure 1 illustrates the diversity we typically find in our translated summaries using summaries of the movie "Day of the Dead". While the Spanish summary mentions the ratio of humans to zombies, this fact is omitted in all other summaries. The French version, the shortest of all, does not mention any of the characters by name, whereas the Italian one already starts introducing them in the second sentence. Additionally, the French version

[2] https://github.com/markriedl/WikiPlots/

A zombie apocalypse has ravaged the entire world. A handful of surviving humans live within a secure underground bunker housing scientists and soldiers in the Everglades. […]

English Original

The dead have awakened and have been ruling the globe for some time now. The film opens with Jamaican pilot John leading a group of people made up of Sarah, […]

Translation from Italian

Following the global invasion of the undead, a small group arrives by helicopter to search for possible survivors. Disappointed, they return to their base, a fortified military camp. […]

Translation from French

Cannibal zombies have taken over the world and now the ratio is one living human for every 400,000 zombies leaving humanity on the brink of extinction. The few remnants of the U.S. government and military are hiding in shelters and small colonies as scientists race against the clock to find a solution to the zombie pandemic. […]

Translation from Spanish

The end of the human race seems to be upon us. The only thing left in the cities are man-eating undead. […]

Translation from German

Figure 1: Opening lines of the summaries we collected for the movie "Day of the Dead".

starts with the opening scene rather than a general description of the world. Generally, we can observe, in these example summaries and across the whole dataset, that the summaries come with varying levels of detail and that each summary focuses on its own set of facts, with a central set of facts that is shared across most summaries. This observation is in line with assumptions made by Nenkova and Passonneau (2004) in assessing summaries. Their approach evaluates an individual summary against an existing set of reference summaries by checking if the summary in question contains facts that are represented in most reference summaries.

An additional observation from qualitative analysis is that some summaries are limited to the premise of the story, whereas others retell the entire story (for an example of a summary that is largely limited to the premise see Figure 5). Some stories also contain meta information, like in this example from a summary of "Willehalm" by Wolfram von Eschenbach: "For we are reminded by the author in Book I […]" which does not strictly focus on the narrative but instead on the reception of the original text.

3.1. Duplicates

Some portion of the summaries can be considered duplicates, being either based on the same external source or one language version being a more-or-less direct translation of another. For training, this may not represent a problem as a paraphrased summary, as produced by back-translation, can be considered the trivial case of a similar story. For the purpose of comparing to previous work, however, we would like to mark these duplicates of sorts as such.

For an initial estimate of the number of duplicates, we annotate 100 pairs of German and English summaries with respect to whether they are direct translations of one another, finding 13 pairs to be direct translations.

To detect these automatically, we experimented with various similarity measures, and after not achieving success with either sentence embeddings or BLEU scores, we settled on BERTScore (Zhang et al., 2020), which yields good accuracy. We use half our annotations to manually tune the decision boundary. This way, we achieve a recall of 0.77 for the non-duplicates and one of 0.83 for the duplicates, both on the held-out half. We provide the similarity scores along with our dataset to allow pruning with a desired decision boundary (we use 0.6 in this work).

4. Dataset Information

Using the methods outlined in Section 3, we collected a dataset across 29,504 movies and books. Most summaries originate from the German Wikipedia, with 28,942 summaries, and the English Wikipedia with 26,385 summaries. This does not accurately reflect the number of overall summaries present in the Wikipedia language versions, as we start from articles that exist in both English and German and have language-specific extraction code. After de-duplication across language versions, our dataset contains 96,831 summaries, for an average of 3.28 summaries per story. Movies are, with 25,860 story-level instances, represented much more frequently than books of which our dataset

Language	Stories Covered		
	All	No Duplicates	Δ
German	98.09%	86.11%	12.00
English	89.43%	89.43%	-
French	74.10%	74.05%	0.05
Italian	66.89%	52.70%	14.19
Spanish	43.25%	29.23%	14.02

Table 1: The number of summaries in the respective languages in our dataset.

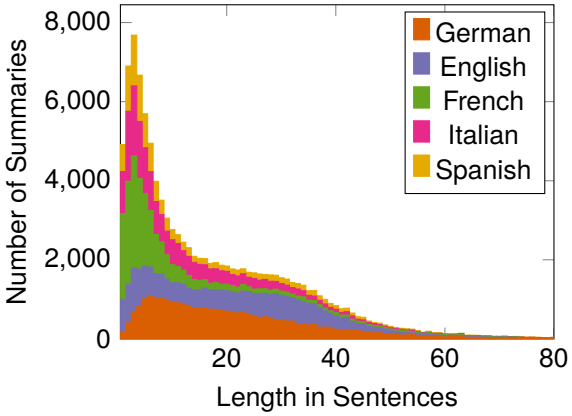

Figure 2: The distribution of lengths of summaries broken down by language before removing direct translations.

only contains 3644 distinct ones[3].

Figure 2 shows the distribution of the summary lengths in sentences in our dataset broken down by language before the removal of exact translations (as the removed pairs cannot be attributed to a single language). We use the sentence splitter by Wicks and Post (2021) with language-specific models where available. The most frequent summary length, in terms of sentences, in our dataset is three. Further, we can observe a bias towards short summaries in the case of the French data. While we suspect that some of this is related to the style of articles across language versions, i.e. generally longer articles in German and English, we can not attribute it to this difference with certainty as it could also be an artifact of our extraction strategies and implementation. While summaries shorter than three sentences may in many cases not be as helpful in the context of story summarizations, we opt to leave such filtering to users of the dataset if desired. The long tail of summaries with a length of 20 or more sentences still includes 40,273 summaries (before de-duplication).

[3]These numbers not adding up to our total is explained by three works being no longer considered either a book or a movie as they were edited during our scraping process and two being tagged as both a book and a movie.

We provide a random 80/10/10 split of our dataset while ensuring that all instances from the dataset by Chaturvedi et al. (2018) are included in our test split. As a result, a model trained on our dataset would not be poisoned for testing on their dataset. As an additional consequence, our test dataset potentially becomes more difficult in that the summaries belonging to the same remake are not labeled as equivalent.

In terms of metadata, the Wikidata link allows for a large variety of information to be retrieved. For example, 344 of our stories are linked to full-text versions by virtue of a Project Gutenberg[4] ID. In terms of books, 476 stories carry associated ISBN information, also opening a path to retrieving their full text automatically. For movies, almost all are linked to multiple movie databases allowing links to further information (e.g. movie reviews).

4.1. Genre Annotations

Genre information can facilitate a wide range of downstream analysis and is included in our dataset due to the extraction from Wikidata. For example, one could validate the capability of embeddings in representing stories by checking if they can be used to distinguish genres. Genre annotations follow a tagging scheme where each story can be associated with multiple genres, the 2013 Spike Jonez movie "Her", for example, is considered a *science fiction film*, a *romance film* as well as a *drama film*.

Figure 3 shows the most frequent genres in our dataset. While most stories are tagged with one or more genres, about 3.7% of stories are not tagged with any genre. Due to the community-annotated nature of the labels, we expect them to not be very consistent (Shenoy et al., 2022) but believe that they could still prove useful for further analysis.

5. Experiments

In our experiments, we attempt, given a summary, to retrieve story summaries for the same source work. In our first experiment, we compare our dataset to the movie remake dataset by Chaturvedi et al. (2018), assessing our dataset's relative difficulty, and in the second we set a baseline for retrieval on our dataset. Both experiments are conducted on the raw summary texts as well as anonymized versions of them where the entities are replaced such that a retrieval based only on token or entity overlap is no longer as effective.

5.1. Anonymizing Entities

In order to maximize performance on our dataset, a good strategy is to look for overlapping named

[4]https://www.gutenberg.org/

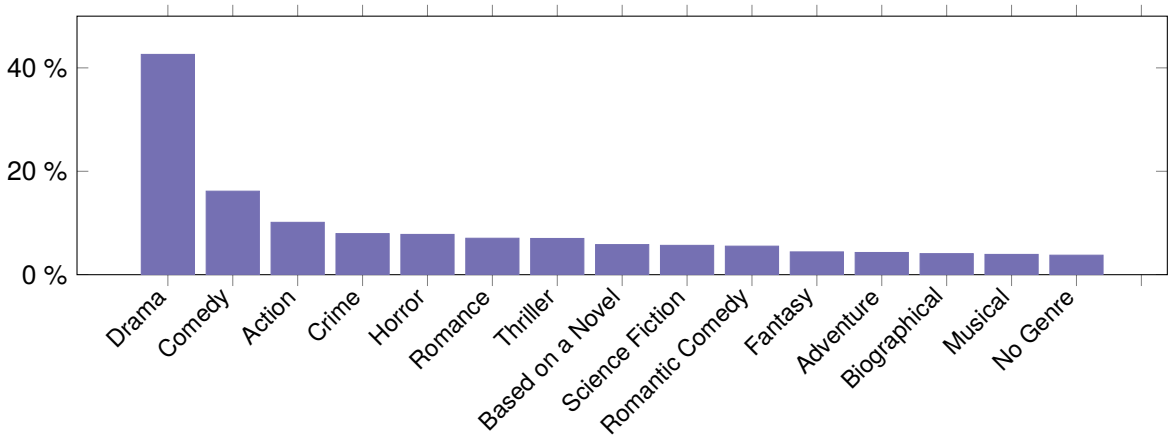

Figure 3: Percentage of stories tagged with each of the top 16 genres.

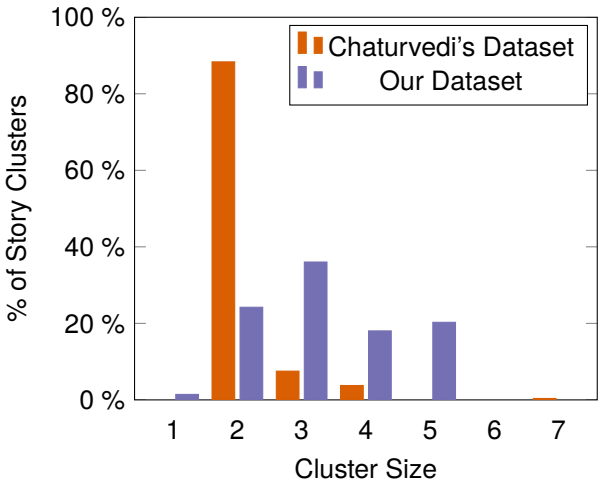

Figure 4: Cluster sizes in our dataset as compared to the one by Chaturvedi et al. (2018). As we limit ourselves to five source languages, none of our clusters are larger than five summaries.

entities, an approach which we consider as a baseline below. This would exploit the fact that different summaries of the same story tend to use the same names. We argue that this is not a desirable strategy in the pursuit of narrative semantics. Thus we experiment with anonymizing entities in the summaries. We follow a best-effort approach to this but acknowledge that there will be a wide range of errors introduced by our replacement process.

Since our dataset is primarily intended as a training dataset the effect of entities being used to recognize story pairs is unintended. We explore an automated method to remedy this effect. On the basis of coreference resolution, we replace all mentions of entities with new names in a consistent manner. We rename every mention of an entity by replacing it with strings of the form "Entity A", "Location B", or "Organization C" depending on the named entity tag. For people, we use an alternative approach, sampling random names associated with the same gender following the name distribution in US-census data.

In terms of implementation, we use Flair's named entity recognition (Akbik et al., 2019) and the coreference resolution model by Xu and Choi (2020) which achieves close to state-of-the-art performance while being computationally much cheaper than the currently best-performing model (Bohnet et al., 2023).

Our approach is limited in the accuracy of the coreference system. Additionally, pre-trained models will have encountered relatively few instances of "Organization C" in their training data. Replacement of entities by strings from the same class (e.g. consistently replacing a country's name with a different one) could yield improvements but can also lead to misleading semantics.

Figure 5 illustrates our anonymization system's capabilities. We found it to generally perform very well on short summaries but deteriorate on longer documents because of missed coreferences. People, or more accurately characters in their respective stories, also appear to be handled better than other entities. Additionally, we replace entire noun phrases when the coreference system returns them as we do not want to rely on descriptions of entities either.

5.2. Detailed Setup

Both experiments are performed using four different sentence encoder models, specifically sentence-T5-large (Ni et al., 2021), all-mpnet-base-v2[5], LaBSE (Feng et al., 2022), and a version of LaBSE fine-tuned for narrative similarity (Hatzel et al., 2023). The rationale is that we want to use a strong sentence model in the form of T5 as well as a widely

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

> The housekeeper Katharina Blum is arrested for spending a night with robber Ludwig Götten, whom she met at a party. […] Meanwhile, the tabloid newspaper Zeitung digs into Katharina's private life, publishing intimate details and falsifying news in order to sell more copies. […] The only tool Katharina has left to defend her lost honor is revenge.
>
> *Original Text*

> Natalie is arrested for spending a night with Edward. […] Meanwhile, the tabloid newspaper Organization U digs into Natalie's private life, publishing intimate details and falsifying news in order to sell more copies. […] The only tool Natalie has left to defend her lost honor is revenge.
>
> *Renamed Entities*

Figure 5: Example of our system anonymizing a story summary of the 1975 movie "The Lost Honour of Katharina Blum". Note that the story's ending is left open as is the case for many shorter summaries in our dataset.

deployed but weaker model in the form of all-mpnet-base-v2, whereas the LaBSE-narrative variant is intended to compare against a model that aims to operate on narrative with the base LaBSE variant serving as a baseline for said model. We use the first few sentences of each text, filling the respective model's context size. This is a practice that clearly has limitations but is established for relatively short texts like ours (e.g. by almost all participants in the shared task by Chen et al., 2022). Retrieval is subsequently performed using cosine-similarity on the models' embeddings. Additionally, we test a baseline system that performs TF-IDF cosine similarity retrieval on just the tokens tagged as belonging to entities by the Flair entity tagger employed in Section 5.1.

For both our experiments we follow Chaturvedi et al. (2018) in reporting the fraction of summaries for which a relevant one (i.e. one of the same story) was retrieved as the most similar one. This metric can be referred to as precision at one.

5.3. Experiment A: Comparative Difficulty of Retrieval in Our Dataset

In our first experiment, in an effort to put subsequent results on our dataset into context, we follow the evaluation setup by Chaturvedi et al. (2018) on their data. That is to say, we retrieve one of the other summaries from a movie's remake cluster by looking for similar summaries. To meaningfully compare our dataset to theirs, we randomly sample the same number of clusters of desired lengths in accordance with the distribution in the movie remake dataset. The only exception to this is one remake cluster of length 7 in the movie remake data, which we instead replace with a summary cluster of length 5, as our dataset contains a maximum of 5 summaries for one work. Accordingly, given that summaries for the same work are likely to be more similar than summaries from the same remake cluster, we expect the performance of retrieval models on our dataset to be much better.

5.4. Experiment B: Baseline Results

The second experiment is intended to build baseline results for our dataset. We operate on our entire test set of 2951 stories with 9718 summaries, setting results for future work to compare against. Further, we provide results when considering only the long summaries, those with 20 or more sentences, in our test set.

6. Results

The best-performing model on the movie remake data (see Table 2) reaches a hit rate of 0.644, marginally outperforming the original story-kernel approach by Chaturvedi et al. (2018), while the best-performing model on our data reaches 0.925. We attribute this difference in performance across datasets mostly to the fact that both stories and entities differ at least slightly across remakes in the same cluster, more so than in different summaries of the same work. For both datasets, we observe a steep drop in performance when anonymizing entities. In principle, there are two reasons for why we may see this drop **(i)** the anonymization works and we see an inability of models to capture narrative semantics with them relying instead mostly on mentioned entities, and **(ii)** our anonymization techniques mangle texts such that the semantics are not preserved. While we expect that for some documents the second reason at least plays a part, we see a promising sign in the comparison of the two LaBSE models. The performance drops steeply in both cases, yet the drop is much smaller for the model fine-tuned for narrative similarity (with an absolute difference of 0.397 as compared to 0.536), indicating that training on the narrative helps the model perform better on the anonymized data. A similar effect can be observed for the sentence-T5 and the mpnet model, leading us to believe that the much larger T5 model is also more capable of capturing narrative semantics. The T5 model outperformed both a sentence model specifically trained for narrative schema similarity (Hatzel et al., 2023) and the original story-kernel approach

	Movie Remakes		Translated Summaries	
Model Name	Text	Anonymized	Text	Anonymized
sentence-T5-large	**0.644**	**0.461**	0.905	**0.686**
all-mpnet-base-v2	0.609	0.300	**0.925**	0.382
LaBSE	0.476	0.204	0.865	0.329
LaBSE-narrative	0.622	0.339	**0.925**	0.528
Entities Bag of Words	0.410	-	0.850	-
Bag of Words Chaturvedi et al. (2018)	0.558	-	-	-
Full Model Chaturvedi et al. (2018)	0.637	-	-	-

Table 2: Precision at one in retrieving summaries from the same cluster on a subset of our data replicating the distribution by Chaturvedi et al. (2018) compared with the performance on their original data.

Model Name	Regular	Anonymized
sentence-T5-large	**0.920**	**0.575**
all-mpnet-base-v2	**0.920**	0.248
LaBSE	0.882	0.228
LaBSE-narrative	0.919	0.405
Entities-BoW	0.877	-

Table 3: Performance of sentence transformer models and our entity baseline, given as precision at one, on our test split. We compare the same models on anonymized (see Section 5.1) and regular versions of the text.

Model Name	Regular	Anonymized
sentence-T5-large	0.665	**0.506**
all-mpnet-base-v2	0.689	0.214
LaBSE	0.638	0.219
LaBSE-narrative	0.680	0.372
Entities-BoW	**0.763**	-

Table 4: Performance of sentence transformer models and our entity baseline, given as precision at one, on a subset of our test split only including summaries that are at least 20 sentences long. This filtering step limits the number of summaries to 3537 from 9848 in the full test set.

by Chaturvedi et al. (2018).

The results of the retrieval models on our dataset are listed in Table 3. We achieved very good results all on the text containing the original entities, with all models delivering a correct response in around 90% of cases. The same trend we encountered on the movie remake dataset, in terms of model performance dropping for the anonymized versions, can be observed on our data.

At first sight, it seems unintuitive that the results in Table 3 are slightly better than the one for the smaller split in Table 2, given that our test split has about twice as many documents. We attribute it to the fact that there are many more summaries of the same story in our test split (3.28 rather than 2.22). At the same time, these results indicate that our choice to include all stories from the movie remake dataset in our test set did not make the retrieval task unduly more difficult due to remakes of the same work not being labeled as identical stories.

The results in Table 4 indicate that retrieval on long documents from our dataset performs much worse, with the best model's performance dropping from 0.920 (see Table 3) to 0.689, now being surpassed by the baseline entity bag-of-word model. We attribute the increased performance of the mpnet and LaBSE-narrative models to the fact that they accept longer input sizes of 384 and 512 sub-word tokens respectively, unlike the LaBSE-base variant and T5, which both only accept 256 tokens. Interestingly, we only observe a slight drop in performance for longer documents in the case of anonymized documents. This may be related to our models not missing out on entity information due to anonymization if said information is cut off due to input-length limitations beforehand.

We observe that the entity bag-of-words baseline achieves a precision at one of 0.410 on the Chaturvedi et al. (2018) dataset whereas it reaches a performance of 0.847 on the data from our dataset. For our data, the results are comparable to the sentence encoder models, while, compared to the same models, the baseline approach severely underperforms on the movie remake dataset. This indicates that entity names in summaries differ across remakes more than they do across the language version in our dataset. When considering only long summaries from our dataset (see Table 4), the entity baseline outperforms all sentence encoder models reaching a precision at one of 0.763 where the best sentence encoder only reaches 0.689. This performance gap further illustrates the strength of entity information as a feature that is, in long documents, partially hidden from the sentence encoders due to truncation.

7. Limitations

The impact of machine translation on downstream models is hard to foresee, but others have found success with using automatically translated texts as training data (Beddiar et al., 2021). After a manual assessment of the translation quality, we are confident that any negative impact on downstream models will be small. Our dataset's largely automatic creation process brings with it the risk of quality issues and Wikidata ontology information is not always consistent or complete. In terms of training similarity models, it is to be noted that while positive sampling should have a close-to-perfect accuracy, negative sampling runs into the issue that remakes and book and movie versions of the same story are classified as distinct stories.

We did not quantitatively evaluate our anonymization approach. It is to be noted that generally using automated data-augmentation techniques can amplify or introduce biases in the data. In our case, this is specifically the case as any error in our coreference resolution system is, in essence, propagated to any similarity model trained on the data. Similarly, our pseudonymous names follow US-census distribution and are thus regionally biased.

All sentence encoder models we employ actually operate on the first few sentences of the summary, truncating text that does not fit the model's input size.

8. Conclusion

In this work, we introduce a novel dataset of story summaries intended as a training dataset for narrative representation tasks. Our dataset is, compared to previous work, of a much larger scale and comes with rich metadata thanks to the Wikidata knowledge graph and can thus be used for a wide range of applications. The dataset, when using the summary texts in their original form, represents a much easier retrieval task than a comparable but smaller-scale dataset by Chaturvedi et al. (2018). Having demonstrated that entity overlap is in fact a very strong narrative-agnostic baseline for the retrieval of matching summaries, we address this by automatically renaming entities, thereby making a name-based matching impossible and thus preparing our texts to be used as training data for narrative-focused rather than entity-focused similarity models. The fact that our renaming or anonymization approach strongly inhibits the retrieval models' performance illustrates their reliance on entities rather than the actual narrative although we do observe narrative-specific and larger models already perform much better than others. This work paves the way for the semantic and narrative modeling of stories based on story summaries by providing a relatively large-scale training dataset. Thanks to splits that are compatible with previous work, we enable comparison with existing datasets after training using our dataset. We envision representation learning approaches on our data enabling the creation of embeddings that can represent narratives.

9. Acknowledgements

The work was supported by the German Research Foundation (DFG) under grant `BI 1544/11-2` as part of the project "Unitizing Plot to Advance Analysis of Narrative Structure (PLANS)".

10. Bibliographical References

Djamila Romaissa Beddiar, Md Saroar Jahan, and Mourad Oussalah. 2021. Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, 24:100153.

Gayatri Bhat, Avneesh Saluja, Melody Dye, and Jan Florjanczyk. 2021. Hierarchical Encoders for Modeling and Interpreting Screenplays. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 1–12, Online. Association for Computational Linguistics.

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference Resolution through a seq2seq Transition-Based System. *Transactions of the Association for Computational Linguistics*, 11:212–226.

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.

Snigdha Chaturvedi, Shashank Srivastava, and Dan Roth. 2018. Where Have I Heard This Story Before? Identifying Narrative Similarity in Movie Remakes. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Volume 2 (Short Papers), pages 673–678, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022. 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.

Mark A. Finlayson. 2012. *Learning Narrative Structure from Annotated Folktales*. Doctoral Thesis, Massachusetts Institute of Technology.

Philip John Gorinski and Mirella Lapata. 2018. What's This Movie About? A Joint Neural Network Architecture for Movie Content Analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1770–1781, New Orleans, Louisiana, USA. Association for Computational Linguistics.

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*, volume 30, pages 2727–2733, Phoenix, Arizona, USA.

Hans Ole Hatzel and Chris Biemann. 2023. Narrative Cloze as a Training Objective: Towards Modeling Stories Using Narrative Chain Embeddings. In *Proceedings of the 5th Workshop on Narrative Understanding*, pages 118–127, Toronto, Canada. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Vladimir Propp. 1968. *Morphology of the Folktale*, 2nd edition. Number 9 in Publications of the American Folklore Society. University of Texas Press, Austin.

Elena Rishes, Stephanie M. Lukin, David K. Elson, and Marilyn A. Walker. 2013. Generating Different Story Tellings from Semantic Representations of Narrative. In *Interactive Storytelling*, Lecture Notes in Computer Science, pages 192–204, Cham. Springer International Publishing.

Kartik Shenoy, Filip Ilievski, Daniel Garijo, Daniel Schwabe, and Pedro Szekely. 2022. A study of the quality of Wikidata. *Journal of Web Semantics*, 72:100679.

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively Summarizing Books with Human Feedback. (arXiv:2109.10862).

11. Language Resource References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Snigdha Chaturvedi, Shashank Srivastava, and Dan Roth. 2018. Where Have I Heard This Story Before? Identifying Narrative Similarity in Movie Remakes. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 673–678, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2022. Model and Data Transfer for Cross-Lingual Sequence Labelling in Zero-Resource Settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hans Ole Hatzel, Fynn Petersen-Frey, Tim Fischer, and Chris Biemann. 2023. Dimensions of similarity: Towards interpretable dimension-based text similarity. In *ECAI 2023*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pages 1004–1011, Krakow, Poland.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BookSum: A Collection of Datasets

for Long-form Narrative Summarization. (arXiv:2105.08209).

Nasrin Mostafazadeh, Lucy Vanderwende, Wen-tau Yih, Pushmeet Kohli, and James Allen. 2016. Story Cloze Evaluator: Vector Space Representation Evaluation by Predicting What Happens Next. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 24–29, Berlin, Germany. Association for Computational Linguistics.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. (arXiv:2108.08877).

NLLB-Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. (arXiv:2207.04672).

Thijs Scheepers. 2017. Improving the compositionality of word embeddings. Master's thesis, Universiteit van Amsterdam, Science Park 904, Amsterdam, Netherlands.

Rachel Wicks and Matt Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3995–4007, 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.

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*, Online.