

Semi-Automated Identification of News Story Chains: A New Dataset and Entity-based Labeling Method

Fatih Gedikli¹, Anne Stockem Novo¹ and Dietmar Jannach²

¹University of Applied Sciences Ruhr West, Mülheim an der Ruhr, Germany

²University of Klagenfurt, Klagenfurt, Austria

Abstract

Automatically deciding if two or more news articles cover the same event—thereby building a *story chain*—is an important problem in news analytics, and knowledge about such story chains can be used, for example, in recommendation scenarios to suggest follow-up news articles. While content analysis on the level of individual news articles and on general news *topics* is well-studied, research on news story chains is still limited, partly due to the difficulty of manually labeling or automatically detecting story chains in larger collections of news articles. In this work, we present a novel (semi-)automated method based on clustering and Named Entity Recognition for creating a dataset for news story chains. An experimental analysis of our method shows that it is highly effective in correctly detecting unrelated stories and identifying candidates for related stories. Thus it helps to reduce manual labeling efforts by 80% without affecting the quality of the dataset. It can even improve the quality of the dataset as manual work is put into only the potentially relevant cases. As an additional result of our work, we publish a new dataset of Business Energy News which was created with the help of our method to foster research in this area.

Keywords

News Story Chains, Follow-up News, Clustering, Datasets, Recommendation

1. Introduction

Automated news analytics methods are widely used in practice today. These methods, which are also well-studied in the academic literature, are for example used to categorize articles according to their topics, to recognize entities that appear in them, to classify them as potential fake news, or for sentiment analysis [1, 2, 3]. What has *not* been studied in similar depth, however, is the automated identification of news *story chains* [4, 5]. Story chains are, roughly speaking, a collection of news articles that report on the same *event*. An event could for example be an earthquake that just happened or a recent plane crash. Note that according to common definitions in the literature, the elements of a story chain are not necessarily follow-up stories, as the name may suggest. They could also be articles on different news outlets that report on the same event simultaneously, maybe also from different angles.

Knowing which articles in a given collection of documents form a story chain can be helpful in different application scenarios. In news recommendation [6], for example, one could use

INRA'21: 9th International Workshop on News Recommendation and Analytics, September 25, 2021, Amsterdam, Netherlands

✉ fatih.gedikli@hs-ruhrwest.de (F. Gedikli); anne.stockem-novo@hs-ruhrwest.de (A. Stockem Novo); dietmar.jannach@aau.at (D. Jannach)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

story chains to recommend follow-up stories to readers who have read an earlier article on an event [7]. Alternatively, one could try to recommend articles from different news outlets that simultaneously report on the same event, e.g., a particular speech by a politician, with the goal of providing a balanced set of viewpoints. Finally, on a news aggregation site, one might use information about story chains to avoid that several articles on the same event are recommended, assuming, for example, that all of them might carry the same (limited) information, e.g., immediately after a catastrophe.

Conceptually, news story chains are located between individual news articles and common news topics (see Figure 1). The degree of abstraction used in news analysis can play an important role, for example in communication sciences, because it may have a significant effect on the results. Think, for example, of a research study on news consumption which counts the number of news items a user is engaged with every day. In such an analysis, the results would change significantly if the study were conducted at the level of individual stories instead of individual articles.



Figure 1: News: Levels of Abstraction

The detection of story chains in a collection of documents is a non-trivial task, and much less explored than common news analytics approaches that focus on individual articles, e.g., in terms of sentiment analysis, or on the identification of general (latent) topics and the automated association of the documents according with these topics. To find out whether two articles deal with the same event and thus are part of an overarching story chain, the articles must be manually compared with each other in pairs. Such comparisons, if made by hand, do not scale to larger collections of documents. Even if we only have $N=100$ documents, which is probably less than what major news sites publish *per day*, $\frac{N \times (N-1)}{2} = 4,950$ such comparisons would be needed.

Besides a certain lack of methods to automate the process of labeling pairs in terms of being part of a story chain or not, not too many public datasets with ground truth information exist that can be used for research on the topic. With this work, we aim to narrow these research gaps. First, we show that the co-occurrence of named entities, e.g., places or people, is a very effective predictor regarding the question if two articles are part of a story chain. We, for example, ran experiments with a clustering algorithm that uses a co-occurrence metric for named entities, finding that the algorithm is highly effective in identifying unrelated articles and in finding candidates for related stories. Second, to foster more research in this area, we provide a new hand-labeled dataset containing articles from the domain of Business Energy News.

The paper is organized as follows. After briefly reviewing previous work in Section 2, we discuss our technical approach to automatically label news articles in Section 3. Afterwards, in Section 4 we present the findings of our empirical evaluation.

2. Background and Related Work

Many works in the literature, as mentioned above, focus either on topic modelling [8, 9] or content analysis on the individual document level [10, 11, 12]. Recently, more attention has been given to the intermediate level where we can find different terms, with a slightly different meaning, like “news events”, “media storms”, or “story chains”.

2.1. Terminology

Topic modeling represents the highest level of abstraction for grouping articles as shown in Figure 1. The latent, underlying theme of an article is modelled usually with unsupervised techniques such as *Latent Dirichlet Allocation* [13] or *Latent Semantic Indexing* [14]. Articles are then grouped by common topics such as, e.g., sports reporting. The term “event” on the other hand is associated with a specific instance of a topic, e.g., the 2020 Olympic Games in Tokyo. Vasterman describes the term “event” as central, covering also the terms “news waves” and “media hypes” [15]. A slightly different term, a “flashpoint”, describes a sudden appearance in the media, usually on brief periods with a rapid fall in interest. Waisbord & Russell discuss the importance of such flashpoints in the frame of today’s digital journalism, which has the power of creating sustained attention on social problems [16].

We follow along with Nicholls & Bright who define story chains as “*normal routines of coverage to dedicate themselves to an exclusive focus on a particular current event, with multiple follow up pieces and different angles explored*” [4]. When an event is tracked in time, the tracking can be further segmented into follow-ups (addressing the same aspect), trees (addressing different aspects), duplicates (rephrasing or correction of previous articles), and summaries (condensing information from a subset of articles). As emphasized by Trilling & van Hoof, a news event can be subject to a “*standalone article that covers one event covered by no one else*”, delimiting it from the definition of a story chain [5].

Furthermore, Nicholls & Bright emphasize the importance of news chain detection when making quantitative analyses of topic coverage in the media [4]. Following their argumentation, articles linked to an event should not receive the same weight in a dataset as the same number of separate articles. Desai & Nagwanshi point out the difference between duplicate and grouping articles [17]. Duplicate articles use the same wording and grouped articles can be syntactically quite different while covering the story from a different angle.

2.2. Data-related Challenges

The difficulties of news story chain detection are mainly twofold: first, news story chains are to some extent underrepresented in general, and thus also in research datasets. “Media storms”, which were characterized by the level of attention and fraction of media coverage, were found to make only 11%, lasting typically for 15 days by Boydston et al. [18]. The database for their analysis is a 10 year study of one US and one Belgian newspaper. This observation was confirmed by Nicholls & Bright on a corpus of almost 40,000 British news articles. They, however, observed a much lower duration of just 1.5 days [4]. While those big stories contain an average of 11 articles, the overall number of story chains with two to five articles was identified

to make a share of 54% of the entire dataset. This is in agreement with a study of three British newspapers, where 30% of the articles were identified as follow-ups of previous articles, i.e., a particular type of news story chains [19].

The second main difficulty is the absence of a time bound. The time span for a story chain is not defined but observed to usually cover a few days [18, 4]. However, related articles can be released months or even years later. Usually a rolling time window of 3 to 7 days is chosen for the determination of document similarity for reasons of computation efficiency [4]. This might lead to an artificially lower number of related articles since a similar event can trigger the mention of a previous news event.

2.3. Story Chain Detection Methods & Their Evaluation

The basis of news story chain detection methods commonly is a quantification of document similarity combined with clustering methods. Clustering algorithms are a common technique for document content analysis [20]. They make use of the distance between documents, trying to minimize the distance of documents inside a cluster (intra-cluster distance) while maximizing the distance between different clusters (inter-cluster distance).

Text documents, which serve as input to the model, can be represented by word frequencies or transformed into word embeddings. Desai & Nagwanshi compared these two representations, a multi-level word embedding approach with TF-IDF models and found that the latter are outperforming the embedding models [17]. They speculate that the variation of entities can be easily handled by the TF-IDF model, while the embedding handles semantic information but diminishes information related to an entity. A combination of both approaches with fine tuning of a BERT model for document similarity and weight adjustment outperforms either of the two models.

Trilling & van Hoof also draw the conclusion that it is advisable combining both methods in order to achieve more robustness, based on a study of 45,000 news articles from different sources [5]. Their investigation shows that using only a TF-IDF approach does not provide enough flexibility for variations in phrasing. Word embeddings overcome this problem at the cost of grouping events which do not form a story chain. It is noteworthy that the authors discovered different event characteristics depending on the news source. Thus, for an unbiased analysis of the content, a variety of news outlets should be considered. Otherwise, the potential of generalization is diminished.

A further aspect on the TF-IDF approach was mentioned by Nicholls & Bright [4]. In order to be computationally efficient, it is a common technique to crop the vocabulary size for the least frequent words. Infrequent words are however often a good indicator for identifying news story chains and thus should be given more importance.

One of the challenges of news story chain detection is the lack of ground truth data for evaluation. Most previous works are based on hand-crafted datasets, e.g., [18, 4]. Such an approach leads to strong control over the labeling process, but is too time consuming on large scale. Therefore, (semi-)automated labeling is desired. For example, the multi-news dataset by Fabbri et al. [21] contains 44,000 clusters of articles, with summaries written by humans. Each cluster contains only 2 to 10 articles. Moreover, from originally almost 40,000 UK news articles in the work of Nicholls & Bright, two independent labelers crafted a dataset of 204 randomly

selected articles [4]. This data shows a strong imbalance: only 59 of 20,706 articles were assigned to story chains. Note that the articles of the dataset were randomly selected and the corresponding article pairs were hand-coded afterwards. Unlike such a random-based approach, we take a data-centric approach and propose an article selection procedure to systematically select articles to improve the quality of the dataset.

In addition to random-based approaches, we also find cluster-based approaches in the literature, where news articles are grouped into story chains using a story chain detection algorithm first. After that, only the article pairs within the groups are labeled ignoring possible relations between the groups.¹ Nicholls & Bright [4], for example, construct an additional dataset for validation by sampling 25 of the story chains generated by their two-step story chain detection method because their random-based development set was quite unbalanced. In the first step, the authors measure pairwise similarity between all articles using the BM25F algorithm. The authors use a sliding 3-day window to reduce the computational complexity of the similarity computation. Second, they employ a network partitioning method to group the articles. In contrast, our procedure, which we describe in detail in Section 3, only uses a Named Entity Recognition (NER) model and is not restricted to a time window at all.

Trilling & van Hoof [5] also construct a dataset by sampling clusters (story chains) covering the same news event.² For each article within an event they manually annotate whether it belongs to the main event, i.e., the event about which the majority of the articles in the cluster are about. Note that Trilling & van Hoof mention that it is hard to check recall but not precision. It is quite straightforward to identify falsely positive assigned articles to story chains. On the other hand, on such vast datasets, it is very difficult to tell if all articles of an event were found. In our work, however, we examine not only the “false positive” rate but also the “false negative” rate of our procedure by comparing all articles in pairs across groups in our new dataset of Business Energy News. Furthermore, we use the validation dataset from Nicholls & Bright to evaluate our approach on data that was not collected by us.

3. Constructing a Novel News Story Chain Dataset

As indicated above, we propose to use the co-occurrence of named entities as indicators that two news articles refer to the same event, i.e., that they form a story chain. Furthermore, as done in previous works, we use a clustering technique to create potential story chains, using a similarity measure that is based on the co-occurrence of named entities. Applying these techniques to a larger collection of news articles ultimately allowed us to create a new dataset, which we use for our evaluation and which we also share for research purposes.

3.1. Underlying Data

As a basis for our research, we first created a large corpus of news stories by automatically harvesting newly published articles from 100 different news sources, including, for example,

¹This approach thus only allows the calculation of the “false positive” rate.

²We could not use their dataset for evaluation because it consists of Dutch news articles and our NER model was trained on English news only.

Reuters³ and ZAWYA⁴. Continuing our previous research in this area, we use an existing processing pipeline and customized web scrapers to retrieve and analyze news from the energy sector. We thus use “Business Energy News” as an example domain⁵ and extract the following data from each article: title, summary, body, date published, and image URL. Between June 6th, 2021 and June 29th, 2021 we collected 2,789 business news articles from the energy sector. From these 2,789 news articles, we automatically determined 100 articles with the article selection procedure we describe below in Procedure 1. We used these 100 articles to populate an unlabeled dataset with potentially many related articles, i.e., we generated all 4,950 pairwise comparisons and used this file as a template for a subsequent manual labeling step. We have implemented our data collection approach in Python. The source code and the fully-labeled dataset are available online.⁶

3.2. Named Entity Recognition Approach

The first step of our processing chain designed to create news story chains from a collection of articles is to detect the named entities in each of them. In our work, we used the Simple Transformers Library⁷ for training a Named Entity Recognition model (NER model). We chose the pretrained transformer based model RoBERTa [22] with 12 layers, 768 hidden nodes, 12 heads, and 125M parameters and fine-tuned the model on the English CoNLL-2003 named entity dataset, which is a collection of news articles from the Reuters Corpus.⁸ The dataset consists of a training file, a development file, and a test file. The NER model was fine-tuned on the training data alone. The development data was used for tuning the parameters of the model. Table 1 shows the final hyperparameters we used for fine-tuning the model.

Table 1
NER Model Hyperparameters

Train batch size:	16
Gradient accumulation steps:	16
Learning rate:	3e-5
Number of train epochs:	15
Max. sequence length:	512
Sliding window:	True

The CoNLL-dataset contains entity tokens for persons, locations, organizations, and names of miscellaneous entities (misc) which do not belong to the previous three groups. Since miscellaneous entities are not suitable for grouping news articles, we have decided to focus only on persons, locations, and organizations in this work.

³<https://www.reuters.com>

⁴<https://www.zawya.com>

⁵Note, however, that our story chain labeling method is not specific to this domain.

⁶https://github.com/fatih-gedikli/news_story_chains-2021-06

⁷<https://simpletransformers.ai>

⁸Data files can be found on <https://www.clips.uantwerpen.be/conll2003/ner/>

3.3. Named Entities Shared Measure (NESM)

In our clustering approach described next we use the Named Entities Shared Measure (NESM) [23]. This measure explicitly takes into account the number of shared entities between two articles and the different entity types:

$$\text{NESM}(d_1, d_2) = \sum_{cat \in \{org, per, loc\}} \frac{\text{NE}(cat, d_1, d_2)_{shared}}{\text{NE}(cat)_{max}} \quad (1)$$

$\text{NE}(cat, d_1, d_2)_{shared}$ is the number of different named entities of a given category cat shared by two documents d_1 and d_2 . $\text{NE}(cat)_{max}$ is the maximum number of different named entities of a given category cat shared by two documents. The NESM-similarity of two documents will range from 0 to the number of entity categories. The range of values is therefore in our case [0, 3].

3.4. Article Selection Procedure

The goal of the article selection procedure is to create a subset of size n of a larger collection of articles in a way that we may assume that many elements in the resulting subset are associated to one story chain. Procedure 1 shows the details of the article selection procedure.

First, the articles of the input dataset are clustered using the NESM-metric and a similarity score cutoff value of 0.4. We use a simple clustering technique in our approach, which creates new clusters whenever an article to be clustered is too dissimilar from existing clusters. Thus, our method does not require a number of clusters to be set in advance.⁹ If an article to process is not too dissimilar, the article is added to the cluster that contains the most similar article found so far. Other clustering techniques from the literature might be applied as well.

Once the clusters are created, we randomly select a cluster with a given minimum cluster size and add random articles from the cluster to the result set, until the maximum number of articles from one cluster is reached or the result set contains n articles. The intuition behind our approach is that the items within a cluster are related with higher probability leading to a dataset with many story chains.

Note that we skip duplicate articles and define an article as a duplicate if there is already a news article in the result set with an 80% word overlap in the title.¹⁰ Furthermore, we excluded all clusters dealing with stock markets as the articles in these clusters were very similar and very likely to be generated automatically.

⁹See the online material for details; all parameters used in our experiment were empirically chosen.

¹⁰Some news sites take the news from other sources and slightly change the original title.

Procedure 1: Article Selection

Input: List of news articles

Output: List of n selected news articles

Parameters: $n \leftarrow 100$, $sim_metric \leftarrow \text{NESM}$, $sim_score_cutoff \leftarrow 0.4$,
 $min_cluster_size \leftarrow 2$, $max_articles_from_cluster \leftarrow 30$

$clusters \leftarrow$ Cluster news articles according to sim_metric and sim_score_cutoff

Shuffle $clusters$ and their elements

$articles \leftarrow \{\}$

foreach $c \in clusters$ **do**

if cluster size of $c \geq min_cluster_size$ **then**

foreach $article \in c$ **do**

if $max_articles_from_cluster$ not reached **and** $article$ is not a duplicate **then**

 Add $article$ to $articles$

if n articles found **then**

return $articles$

end

end

end

end

end

return $articles$

4. Evaluation

Remember that the basic assumption in our approach is that every article in a news story chain shares at least one common named entity with all other articles in that story chain, be it a person, a location, or an organization. Thus, even when articles might be very similar in other regards, e.g., in terms of their text embeddings, we consider them to not be part of the same story chain, even though they might be on the same general topic.

We evaluate the appropriateness of this assumption using the following general methodology. We tasked two human judges with a labeling exercise, where the task was to label pairs of news messages. For each pair, the judges (“coders”) had to report if the two news articles were part of a story chain or not. One coder (Coder 1) was one of the authors of this work. The other one (Coder 2) a student who was informed about the concept of a story chain before accomplishing the task. The student was however not made aware of the background or the goals of the study. The outcome of these labeling exercises forms the basis for the assessment of our approach. We provide further details of our evaluation procedure below.

4.1. Datasets

We used two datasets for our evaluation and Table 2 shows their characteristics. The characteristics of the two datasets are quite different, allowing us to gauge the generalizability of our findings at least to a certain extent.

The Business Energy News dataset, discussed in the previous section, is very focused and contains only business news articles from the energy sector scraped from rather unknown but special news outlets such as ZAWYA, pv magazine, and Chemical Engineering. The dataset contains 100 articles resulting in 4,950 article pairs of which 88 (1.77%) are related, i.e., they are part of a story chain according to a human labeler.

The dataset from Nicholls & Bright [4], in contrast, includes news articles for the public taken from BBC News, The Mail Online, The Express, The Guardian, and The Mirror. We used the hand-coded validation data from Nicholls & Bright which was mainly used for the overall evaluation of their method for identifying news story chains.¹¹ The dataset contains 254 articles resulting in 32,131 different article pairs of which a human has classified 126 (0.39%) as related. Using this second dataset also helps us to see how our approach works on data that was not collected by us.

Table 2

Characteristics of Evaluation Datasets

	Business Energy News dataset	Nicholls & Bright dataset
Number of news outlets:	18	5
Number of articles:	100	254
Number of article pairs:	4,950	32,131

4.2. Validation on Business Energy News dataset

To gauge the usefulness of our approach, we gave the two human judges slightly different tasks:

- Coder 1 was working on the assumption that only articles that share at least one named entity are candidates to be part of a story chain, i.e., “related”. Therefore, the coder was presented only with the 901 (18.2%) pairs of articles from the dataset where this was the case. The other $4,950 - 901 = 4,049$ pairs were automatically considered “unrelated”.
- Coder 2, in contrast, was tasked to label all 4,950 pairs of the dataset as being related or unrelated.

This experiment configuration first of all lets us assess how many story chains Coder 1, who relies that the pairs marked as unrelated by our approach are truly unrelated, would miss. Remember that we do *not* assume that all pairs with a shared named entity are definitely part of a story chain, which is why Coder 1 is tasked to label them. However, the number of candidates for related articles pairs, which must be examined manually, is much smaller, i.e., only 901 pairs.

Overall, our method labeled 4,049 of the 4,950 pairs as being unrelated. To determine the rate of “false negatives” (missed chains) in this set, we compared these 4,049 pairs of articles with the results of the human evaluator (Coder 2) and found an accuracy of 99.9%. There was a discrepancy only in three cases that we investigated further. The follow-up analysis revealed that while the pairs of articles were actually similar, they referred to different events and therefore should have been classified by Coder 2 as unrelated, which corresponds to the classification that was made by our procedure. Overall, we can conclude that our heuristic

¹¹Data file “story_pairs_validation_august.csv” can be found on <http://dx.doi.org/10.5287/bodleian:R5qdeJxYA>

based on the co-occurrence of named entities is highly effective in terms of identifying pairs that are definitely *not* part of a story chain.

Next, we examined how many related pairs of news articles were identified by the two coders within the set of 901 pairs that were considered by our method to be potentially related based on the co-occurrence of named entities. After analyzing the cases in which the coders disagreed in more depth—we provide details later in this section—we found that 88 of the 901 pairs (9.76%, total 1.77%) can be considered being part of a story chain. Comparing this rate of identified story chains to the one observed for the Nicholls & Bright dataset, where only 0.39% of the articles were related, indicates that our procedure for creating the research dataset as described in Section 3 was effective. As a result, our final dataset is much less sparse than the one by Nicholls & Bright, making it more suitable to design new methods for news story chain detection.¹²

Let us now take a closer look at coder agreement and disagreement. We first compared the labeling results of both coders. Overall, our analysis showed that the coders agreed in as many as 99.4% of the cases, i.e., on 4,923 of 4,950 article pairs, leading to a Krippendorff’s alpha of 0.87 and strong inter-rater reliability [24]. Table 3 shows the number of article pairs on which both coders agreed (numbers in bold) or disagreed (non-bold numbers). The numbers in parentheses represent the number of article pairs that were automatically classified as unrelated by our procedure used by Coder 1.

Table 3

A confusion matrix visualizing the number of article pairs on which both coders agreed (numbers in bold) or disagreed (non-bold numbers). Numbers in parentheses represent the number of article pairs that were automatically classified as unrelated by our procedure used by Coder 1.

		Coder 2 (unaware)	
		unrelated	related
Coder 1 (semi-automated)	unrelated	4,858 (4,046)	4 (3)
	related	23	65

Overall, there are 27 (= 23 + 4) article pairs where the coders disagreed. First, we looked in detail at the 4 article pairs which were according to Coder 1 unrelated and Coder 2 related (upper right corner of the matrix). Note that an error in this corner would be more serious than an error in the lower left corner because an article pair could be marked as unrelated automatically by our procedure and Coder 1 would have no chance to examine this news pair. We have already discussed 3 of the 4 cases above. The last case was labeled by Coder 1 as unrelated, even though our procedure has detected two shared named entities. Specifically,

¹²We note that to a certain extent the higher density of story chains in our dataset may also be attributed to the fact that our news sources focus on a defined domain, that of energy news.

one of the articles was about the G7 Summit, mentioning the Australian Prime Minister, and the other one about a project launch of a \$43 million program aimed at identifying how to reduce emissions in industry, also in Australia. Since we have defined news chains based on their relation to the same event, this article pair must therefore be labeled as unrelated, and the label by Coder 2 is considered a mistake. Note that Nicholls & Bright have identified this error type as one of the main errors where “*the line between related and unrelated articles becomes somewhat blurry*” [4].

Next we examined in detail the 23 article pairs which were related according to Coder 1 and unrelated according to Coder 2 (lower left corner of the matrix). A detailed description of the errors for each article pair can be found in the online appendix.¹³ Here, we would like to address one interesting error, which was the cause of 4 misclassifications by Coder 2. Specifically, there was one overview article which addressed multiple independent news topics at once.¹⁴ For this article, Coder 2 has apparently overlooked the relationship between a sub-story in this article and other articles that were about the same event. Generally, one can try to avoid such mistakes by breaking overview articles into atomic news in advance or filtering them out. Because of the many shared named entities in the multi-topic article, Coder 1 was in this case alerted and examined the relationship of the article with other articles with increased attention.

Overall the results are in line with our expectations, i.e., news stories in a story chain share named entities and focusing on these alone when hand-coding the dataset does not lead to a loss of precision and may even improve the quality of the resulting data.

4.3. Validation on Nicholls & Bright dataset

We used the validation dataset of Nicholls & Bright to test our approach on unseen news from other news outlets and non-business topics. Figure 2 shows that by considering only article pairs with at least one common shared entity reduces the number of article pairs to be manually examined to 18.2% for the Business Energy News dataset and 17.9% for the Nicholls & Bright dataset respectively.

Next, we examined the proportion of “false negatives”, i.e., the proportion of article pairs which we would incorrectly flag as unrelated because they do not share any named entities. There were only three false negative classifications of article pairs in the Nicholls & Bright dataset. These articles all have in common that they mainly contain video content and only have a very short textual description. Furthermore, the article texts file provided by Nicholls & Bright, which contains the extracted texts of the news articles, does not contain any textual content for these video news. Therefore, no common entities could be identified. If we ignore the video news, a saving of more than 80% in the manual labeling task can be achieved with no error at all.

Finally we compared the relative number of positive examples in both datasets. As expected, due to our article selection procedure we had 4.5 times more positive examples for story chains compared to the dataset from Nicholls & Bright.

¹³See file “diff_coder1_coder2.txt” in the attachment.

¹⁴<https://www.pv-magazine.com/2021/06/22/the-hydrogen-stream-off-grid-hydrogen-power-solution-based-on-alkaline-fuel-cell-from-israel-first-green-hydrogen-production-in-russia>

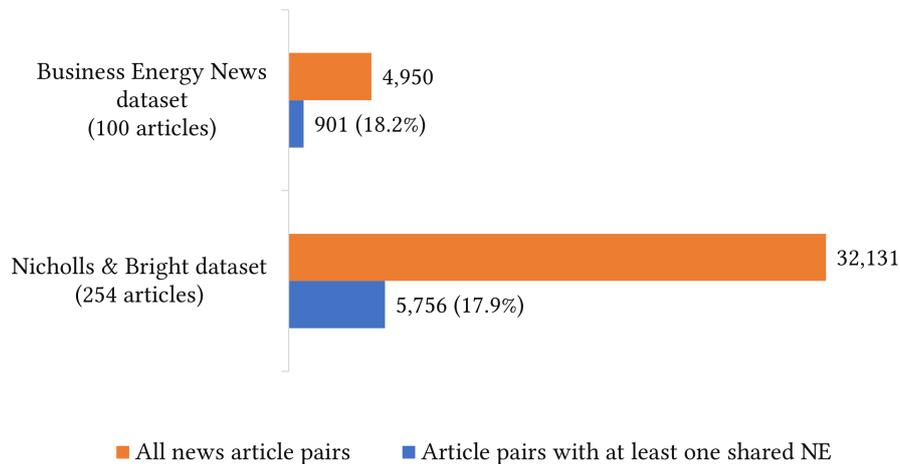


Figure 2: The number of article pairs to be examined with and without NER.

5. Discussion and Conclusion

News story chain detection plays an increasingly important role in news analytics, recommender systems, and also in communication sciences in general. One of the main challenges in training and testing a corresponding model is the quadratic effort in coding the data. Hence, we followed a data-driven approach in this work to counteract the quadratic effort required to manually label a news story chain dataset. Our data collection procedure reduces the manual labeling effort while the quality of the data remains the same. It can even improve the quality of the data in the labeling process as more time can be invested in the potentially relevant cases. Another challenge is the small number of story chains in big datasets. We showed that our article selection procedure can increase the proportion of positive examples.

We tested our approach on a new dataset on business energy news which we offer for use by the wider research community.¹⁵ The results show that with about 80% less coding effort we could detect all positive examples for story chains. Even on Nicholls & Bright’s comparative dataset [4], the coding effort would be reduced by more than 80%. Furthermore, our approach ensures through a smart article selection procedure that more positive examples for story chains can be detected in the final dataset.

In our future work we would like to go beyond Named Entity Recognition by using Named Entity Disambiguation in order to map entity strings to real world entities. This would move us from an exact entity search to a more flexible semantic search which would also allow synonyms to be recognized.

Generally, most research up to this point has focused on improving the accuracy of models and held the data for training the models fixed. We think that a data-centric or data-driven approach offers much more potential and will pave the way for better results in the future.

¹⁵The source code and the fully-labeled dataset are available at https://github.com/fatih-gedikli/news_story_chains-2021-06.

Acknowledgments

We would like to thank Selin Kartal for supporting us with the manual labeling task as part of her bachelor thesis.

References

- [1] D. Billsus, M. J. Pazzani, User modeling for adaptive news access, *User Modeling and User-Adapted Interaction* 10 (2000) 147–180. doi:10.1023/A:1026501525781.
- [2] E. F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, in: *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003, CONLL '03, 2003*, p. 142–147. doi:10.3115/1119176.1119195.
- [3] M. Vicari, M. Gaspari, Analysis of news sentiments using natural language processing and deep learning, *AI & SOCIETY* forthcoming (2020) 1–7. doi:10.1007/s00146-020-01111-x.
- [4] T. Nicholls, J. Bright, Understanding news story chains using information retrieval and network clustering techniques, *Communication Methods and Measures* 13 (2019) 43–59. doi:10.1080/19312458.2018.1536972.
- [5] D. Trilling, M. van Hoof, Between article and topic: News events as level of analysis and their computational identification, *Digital Journalism* 8 (2020) 1317–1337. doi:10.1080/21670811.2020.1839352.
- [6] M. Karimi, D. Jannach, M. Jugovac, News recommender systems - survey and roads ahead, *Information Processing and Management* 54 (2018) 1203–1227. doi:10.1016/j.ipm.2018.04.008.
- [7] W. Gu, S. Dong, M. Chen, Personalized news recommendation based on articles chain building, *Neural Computing and Applications* 27 (2016) 1263–1272. doi:10.1007/s00521-015-1932-x.
- [8] C. Bouras, V. Tsogkas, A clustering technique for news articles using wordnet, *Knowledge-Based Systems* 36 (2012) 115–128. doi:10.1016/j.knosys.2012.06.015.
- [9] T. Basu, C. Murthy, A similarity assessment technique for effective grouping of documents, *Information Sciences* 311 (2015) 149–162. doi:10.1016/j.ins.2015.03.038.
- [10] A. Damstra, R. Vliegthart, (Un)covering the Economic Crisis?, *Journalism Studies* 19 (2018) 983–1003. doi:10.1080/1461670X.2016.1246377.
- [11] A. Leupold, U. Klinger, O. Jarren, Imagining the city, *Journalism Studies* 19 (2018) 960–982. doi:10.1080/1461670X.2016.1245111.
- [12] G. Nygren, M. Glowacki, J. Hök, I. Kiria, D. Orlova, D. Taradai, Journalism in the crossfire, *Journalism Studies* 19 (2018) 1059–1078. doi:10.1080/1461670X.2016.1251332.
- [13] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, *Journal of Machine Learning Research* 3 (2003) 993–1022.
- [14] T. Hofmann, Probabilistic latent semantic indexing, in: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99, 1999*, p. 50–57. doi:10.1145/312624.312649.

- [15] P. L. Vasterman, Media-hype: Self-reinforcing news waves, journalistic standards and the construction of social problems, *European Journal of Communication* 20 (2005) 508–530. doi:10.1177/0267323105058254.
- [16] S. Waisbord, A. Russell, News flashpoints: Networked journalism and waves of coverage of social problems, *Journalism & Mass Communication Quarterly* 97 (2020) 376–392. doi:10.1177/1077699020917116.
- [17] A. Desai, P. Nagwanshi, Grouping news events using semantic representations of hierarchical elements of articles and named entities, in: 3rd International Conference on Algorithms, Computing and Artificial Intelligence, ACAI '20, 2020, pp. 1–6. doi:10.1145/3446132.3446399.
- [18] A. E. Boydston, A. Hardy, S. Walgrave, Two faces of media attention: Media storm versus non-storm coverage, *Political Communication* 31 (2014) 509–531. doi:10.1080/10584609.2013.875967.
- [19] T. Harcup, D. O'Neill, What Is News? Galtung and Ruge revisited, *Journalism Studies* 2 (2001) 261–280. doi:10.1080/14616700118449.
- [20] C. C. Aggarwal, C. Zhai, *A Survey of Text Clustering Algorithms*, Springer US, Boston, MA, 2012, pp. 77–128. doi:10.1007/978-1-4614-3223-4_4.
- [21] A. Fabbri, I. Li, T. She, S. Li, D. Radev, Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1074–1084. doi:10.18653/v1/P19-1102.
- [22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, *CoRR abs/1907.11692* (2019). arXiv:1907.11692.
- [23] S. Montalvo Herranz, V. Fresno Fernández, R. Martínez Unanue, NESM: a Named Entity based Proximity Measure for Multilingual News Clustering, *Revistas - Procesamiento del Lenguaje Natural* 48 (2012).
- [24] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology* (second edition), Sage Publications, 2004.