Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2413–2429
Marseille, 20-25 June 2022

Story Trees: Representing Documents using Topological Persistence

Pantea Haghighatkhah[♢], **Antske Fokkens**[♣♢], **Pia Sommerauer**[♣],
Bettina Speckmann[♢], **Kevin Verbeek**[♢]
♢ Eindhoven University of Technology, Dep. of Mathematics & Computer Science
♣ Vrije Universiteit Amsterdam, Computational Linguistics and Text Mining Lab
{p.haghighatkhah, b.speckmann, k.a.b.verbeek}@tue.nl, {antske.fokkens, pia.sommerauer}@vu.nl

Abstract

Topological Data Analysis (TDA) focuses on the inherent shape of (spatial) data. As such, it may provide useful methods to explore spatial representations of linguistic data (embeddings) which have become central in NLP. In this paper we aim to introduce TDA to researchers in language technology. We use TDA to represent document structure as so-called story trees. Story trees are hierarchical representations created from semantic vector representations of sentences via persistent homology. They can be used to identify and clearly visualize prominent components of a story line. We showcase their potential by using story trees to create extractive summaries for news stories.

Keywords: Topical Data Analysis, Semantic Vectors, Document level discourse

1. Introduction

Topological data analysis (TDA) provides insights into the shape of (spatial) data. TDA is built on a strong mathematical foundation and thus delivers transparent and explainable results. Rather than analyzing data point by point, TDA captures the overall shape of the data, ignoring the inherent noise present in individual data points. As such, TDA can represent point data in terms of global geometric or topological structures that capture information about the data as a whole. It has received significant attention in a variety of scientific fields, such as biomedicine (Nielson et al., 2015), chemistry (Lee et al., 2017), and material science (Nakamura et al., 2015).

Spatial data has become increasingly central in NLP, as most approaches rely on embedding representations of linguistic units. Hence previous work has explored the possibilities of TDA for text analysis in various ways. For example, several papers use TDA to derive superior features from high-dimensional data that can be used as input for machine learning, with some reporting good results compared to baselines (Doshi and Zadrozny, 2018; Kushnareva et al., 2021). Another set of papers uses TDA directly to explore a text's structure for the purpose of clustering or finding key phrases. However, to the best of our knowledge, these attempts use only basic lexical overlap as input for their analyses. In general the use of TDA so far remains limited to isolated attempts; the NLP community as a whole is still largely unaware of the advanced, transparent analysis methods that TDA can provide.

In this paper, we illustrate how to apply TDA to high-dimensional linguistic data representations, while providing a transparent model for investigating a text's structure and salient components. To this end we introduce story trees which capture the shape of a text via the persistence of the connected components of its sentences, represented as points in a semantic space. By tracking the growth process of these connected components we define a hierarchical tree structure that can be used to (visually) identify salient components of the text. Though story trees can be created using any semantic representation (such as sentence-BERT (Reimers and Gurevych, 2019)), we already achieve meaningful results on extracting the most relevant sentences from news articles with basic bag-of-word representations created from word2vec embeddings (Mikolov et al., 2013a).

The remainder of this paper is structured as follows. In Section 2, we first give an intuitive introduction to persistent homology, one of the core methods of TDA. We also summarize previous work using TDA in NLP. Section 3 describes how we create merge trees from sentence representations and trim them into story trees. Furthermore, we introduce our visual drawing style for story trees. In Section 4 we first explain how to create extractive summaries from story trees and then evaluate the quality of the summaries constructed with our methods both in an automated fashion against baselines and via human annotation. This is followed by our discussion and conclusion.

By providing an intuitive explanation of persistent homology and demonstrating how it can be used to approach a realistic NLP task, we aim to illustrate the potential of TDA for transparent analysis of spatial NLP data and provide the basis for further explorations.

2. Background and Related Work

Topological data analysis (TDA) provides methods to study the shape and structure of data. In Section 2.1 we introduce those TDA concepts and methods that are used in this paper and in prior work, in particular, homology and persistence. In our description, we follow Munch (2017) closely, but limit ourselves to those components that are relevant for our study; we also adapt

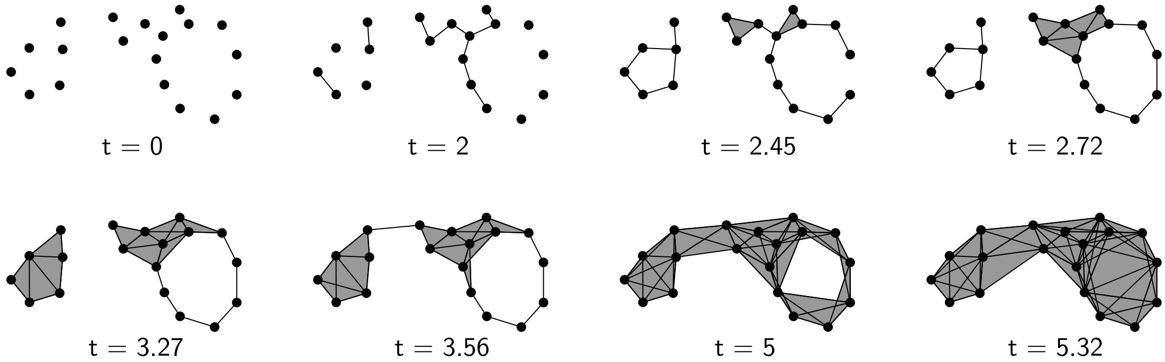

Figure 1: The simplicial complex $G(t)$ for increasing values of t; observe how a set of isolated points at $t = 0$ gradually grows into a single connected component with one hole at $t = 3.56$; at $t = 5.32$ the hole is filled.

the presentation for an NLP audience. In Section 2.2 we discuss related work that makes use of various forms of topological data analysis for text analysis.

2.1. Persistent Homology

In the following we describe persistent homology for spatial data: sets of points in (high-dimensional) space. The first step of a TDA pipeline creates a **topological signature** of the data. A topological signature is a simplified representation of the points that captures their overall shape; it abstracts away from the values of individual data points. A topological signature generally contains (significantly) less information than the full point set. It is hence of paramount importance for reliable data analysis that the topological signature captures the structure and the shape of the data well. In this paper we use **persistence diagrams** as topological signatures; they are created using **persistent homology**.

Homology captures the overall shape of spatial data by considering the holes in the data (in varying dimensionality). A set of points, of course, does not have any holes or even connected components. To be able to speak about such concepts, we need to build a structure on the points. A common such structure is a so-called *simplicial complex*, which we can build as follows. First of all, we need a means to measure the distance between points, such as the Euclidean or the cosine distance. For a given distance value t we can then build a graph $G(t)$ – the simplicial complex – on the points. Each point becomes a 0-dimensional vertex of $G(t)$. We connect two vertices in $G(t)$ by a 1-dimensional edge if the corresponding points x and y are less than distance t apart, that is, $d(x, y) \leq t$. The graph $G(t)$ is a 1-dimensional skeleton of the data. We can now speak about the connected components in the data, by speaking about the connected components of $G(t)$. We can generalize this concept from edges to higher-dimensional building blocks: 2-dimensional triangles connecting three vertices of $G(t)$, 3-dimensional pyramids connecting four vertices of $G(t)$, etc. See Figure 1: triangles are shaded in gray, higher-dimensional building blocks are not indicated for ease of visual interpretation.

The connectivity of the simplicial complex $G(t)$ depends on the choice of t. For $t = 0$ all points are isolated in $G(0)$. As t grows, points are connected and holes are created and eventually filled (see Figure 1). Any specific value of t shows us a particular picture of the data. However, there is not a single value of t that is the "best" or most representative. Rather than picking a specific t, we hence examine how connectivity and holes evolve over a range of values for t. In particular, we are interested in structures that persist over large ranges of t, that is, **persistent homology**.

Imagine that we grow a ball of radius r around each data point. When two such balls touch, then the corresponding points are at distance $t = 2r$. All structural changes in the simplicial complex $G(t)$ happen at such an event of two touching balls: (i) two connected components merge, (ii) a hole appears, (iii) a hole is filled. The value of t when a connected component or a hole appear marks the *birth* of that structure. Correspondingly, we speak of the *death* of a hole when it is filled and hence disappears. Connected components cannot disappear, but they can merge; at this point an arbitrary one of the two dies while the other persists. If applicable, the "older" component (smaller value t at birth) persists. See again Figure 1: at $t = 0$ we have no connections. As we grow t connections start appearing ($t = 2$). When t is large enough ($t = 2.45$) the first hole is born. At $t = 2.72$ the second hole is born. Growing t further results in the death of the first hole at $t = 3.27$ and a merge of the two remaining connected components at $t = 3.56$. Eventually at $t = 5.32$ the second hole dies.

We can now study the persistence of connected components and holes, that is, the time between their birth and death. The intuition is that structures with higher persistence (longer interval between birth and death) are more important for the global shape of the data; structures with low persistence are essentially noise. We use a **persistence diagram** to visualize the persistence of each structure, see Figure 2. The connected components form the so-called 0^{th} homology group H_0 and are indicated in blue. Each connected component corresponds to a blue point in the persistence diagram; the coordinates are determined by the value of t at birth (x-coordinate) and at death (y-coordinate). Similarly, the holes form

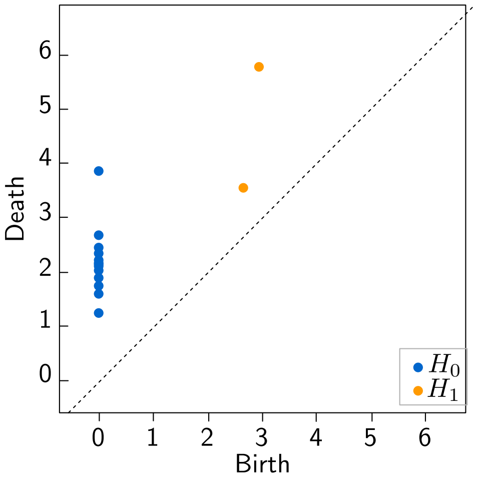

Figure 2: Persistence diagram for Figure 1: blue points are connected components, yellow points are holes.

the 1^{st} homology group H_1 and are indicated by yellow points. The shorter the lifespan of a component, the closer it is to the diagonal; components which are further away from the diagonal have higher persistence.

2.2. TDA in NLP

One of the major advances TDA can bring to NLP is a transparent analysis of rich, spatial representations. To the best of our knowledge, none of the existing approaches achieves this; they are either limited to representations based on lexical overlap or are not transparent. Nevertheless, they provide evidence that TDA has the potential to extract meaningful information from textual data. We distinguish two general directions in which TDA has been used in NLP: (1) using persistence information as features in classification, and (2) using persistent homology directly to reason over data.

Classification. Various approaches use features created using TDA as input for a classifier to perform text classification. The underlying assumption is that the structure of the data captures information about the content of a document based on which different types of texts can be distinguished. For example, Doshi and Zadrozny (2018) achieved state-of-the-art performance in movie genre classification based on reviews. Kushnareva et al. (2021) outperform neural-based baselines with features extracted from BERT using TDA to detect artificially generated texts. Savle et al. (2019) use topological structures to support inference on a complex legal data set. Almgren et al. (2017) cluster images according to popularity based on the cosine similarity of their captions represented by word embeddings. They show that for this task TDA outperforms traditional clustering approaches. Gholizadeh et al. (2018) create graphs based on co-occurring entities in novels and show the resulting topological structures can be used to predict the novel's author. Michel et al. (2017) show promising first steps, that at this point cannot compete with traditional approaches. They use features created using TDA for clustering and sentiment analysis, but report a basic model using tfidf-based features outperforms theirs. Though a more thorough analysis would be needed to reveal why TDA did not yield better results, we suspect that this failure was caused by choices in the setup (loss of too much relevant information) and the metric chosen in their study. The overall outcome of their work nevertheless illustrates that it is not trivial to use TDA in a meaningful way.

The success of some classification-based approaches indicated that TDA can highlight relevant linguistic features. However, these approaches do not provide insights into what type of information TDA highlights and how it informs the classification decisions.

Direct use of TDA. Approaches that are more relevant for our work use persistence information directly to reason over textual data. The intuition behind such approaches is that persistence can directly represent relevant information and thus offer a transparent approach to NLP tasks. To the best of our knowledge, there is no existing approach that uses persistent homology to reason over components of texts represented by high density word embeddings. Previous work limited to lexical overlap does, however, illustrate the potential of such approaches: Chiang (2007) proposes to use TDA features to perform hierarchical document clustering using co-occurrence features as well as features extracted from ontologies to represent documents.

Zhu (2013) uses a derivative of persistent homology to represent the structure of stories and essays based on connections between sentences. The intuition is that repetitions in the texts form connected components. Instead of starting with isolated points (representing sentences), the sequential structure of the text is already imposed; sentences preceding each other are automatically connected (regardless of their distance). Holes are created by similar sentences that do not immediately follow each other. Similarity is defined as lexical overlap between sentences. The structure that arises is a consequence both of lexical similarity and order, and hence not a proper persistence diagram. Zadrozny (2021) combine persistent homology with concepts from network analysis and aim to find circular arguments in a small set of carefully constructed toy data. Christianson et al. (2020) extract semantic networks using co-occurrences of word surface forms from text books.

Closest to our work are Guan et al. (2016) who extract key phrases for summarization from scientific writing based on similarity measured by overlap of phrases. Their approach requires more advanced preprocessing (including syntactic analysis for key word selection), whereas we apply our approach to basic bag-of-word representations of the full sentences in the text. In addition to the differences in approach, our work cannot be compared to theirs directly: our approach relies on the typical story line of news stories which differs from the structure used in scientific writing. Furthermore, they focus on key phrase extraction whereas we focus on sentences.

Overall, the variety of approaches utilizing TDA shows that there is a definitive potential for this line of work. However, TDA cannot be simply used as a "black box"; successful methods make careful choices when modeling an NLP task in topological terms and transparently evaluate the results. Our paper contributes to this line of work: our story trees depend directly on the shape of the text we are analyzing and our proposed drawing style for story trees provides an insightful visual summary of text structure.

2.3. Representing documents through graphs

In this paper, we use extractive summarization as a use-case of story trees. This task has a tradition of being approached from the perspective discourse representations in the form of graph structures. Early implementations (Marcu, 1999) rely on automatic representations of discourse structure following Rhetorical Structure Theory (Mann and Thompson, 1988). In such approaches, discourse parsers are used to identify discourse units (clauses or sentences) and their relations to one another. Discourse units can either express central information (called 'nuclei') or provide additional, non-essential information (called 'satellites'). A summary can be created by selecting nuclei units that are most central to the document structure.

Later approaches rely on clustering sentences and using graph reasoning methods to detect the most central sentence of a cluster or document graph on the basis of semantic distances (Mihalcea and Tarau, 2004; Erkan and Radev, 2004; Zheng and Lapata, 2019). For example, Zheng and Lapata (2019) use Bert sentence representations to build a graph structure of a document. In contrast to earlier approaches, the edges in the graph are directed on the basis of relative sentences location in the document (earlier sentences are assumed to be more important than later sentences). This unsupervised approach outperforms supervised approaches and the strong LEAD baseline (i.e. selecting the first n sentences of a document).

The document representation and summarization approach we present in this paper has similarities to such graph reasoning approaches. We use topological data analysis to build a document graph (i.e. story tree) and determine the sentences that express the most central ideas of a document. We do not claim that our approach is superior to existing, unsupervised graph-based reasoning approaches. Rather, we aim to illustrate the potential of viewing NLP data through the lens of TDA.

3. Story Trees

We propose a model which captures text structure based on persistent homology. In particular, we focus on the 0^{th} homology group: the connected components. To employ persistent homology as described in Section 2.1 we need two ingredients: the text data represented by points in high-dimensional space and a distance function. Our approach can be used with any word embedding on sentence level in a semantic space. For our experiments we use the average of pre-trained word2vec embeddings of the words in each sentence, see Section 4.2 for details. The vectors resulting from the sentence embedding directly correspond to points and these are the points we use for our approach. As distance function between the points (vectors) we use the cosine distance between the vectors (points).

Given the point representation of the sentences in the input text, we now grow balls around each point as described in Section 2.1; we observe the birth and death of connected components. Initially all sentences (points) form individual isolated components; as the distance t increases they gradually grow into a single connected component.

Merge trees. We can capture the growth process of the connected components in a **merge tree**. Each point (sentence) corresponds to a leaf of the merge tree MT. Whenever two components merge, we add an interior node to MT which stores the value t at which the merge occurred; the root of MT corresponds to the final connected component.

Story trees. A merge tree is a direct representation of the evolution of the connected components of our input points (sentences). We propose to trim MT and augment it with additional information to create what we call a **story tree**: a hierarchical representation of the salient parts of a text.

We proceed as follows (see Figure 3 for reference). First of all, we identify all **primary pairs** of leaves in MT. A primary pair of leaves are two leaves of MT (that is, two sentences represented as points/vectors) which merge and form a connected component of size 2. These two sentences are semantically close as measured via the cosine distance. In our example, there are three primary pairs: $(0, 8)$, $(1, 6)$, and $(3, 4)$. In our drawing of the story tree, sentences are represented by labeled squares and are ordered along the x-axis. That is, their x-coordinate corresponds to the order in which they appear in the story we are analyzing. The y-coordinate corresponds to the order of the t values of the first merge a sentence is involved in. Hence, the two sentences that form a primary pair are drawn at the same vertical height; furthermore, the primary pair $(0, 8)$ appears before the primary pair $(1, 6)$, which in turn appears before the primary pair $(3, 4)$.

We next identify all **secondary leaves**: leaves that merge with a primary pair. In our example there are two secondary leaves: 9 which merges with the primary pair $(0, 8)$ and 5 which merges with the primary pair $(1, 6)$. In our drawing we grow the story tree vertically above the left-most sentence of each primary pair. We use the elder rule when determining persistence: the older component (lower t value) is the one that survives the merge. Hence, when the secondary leave 9 merges with the primary pair $(0, 8)$, the component of $(0, 8)$ (which now also contains 9) persists. The length of each

[0] The Glasgow summit presents the climate crisis as a global crisis, masking the fact that rich developed countries are the main culprits and poor countries get much of the burden.

[1] The main problem: citizens of these countries have to change their consumption and lifestyle to help reduce the greenhouse gas emission.

[2] World leaders are worried that their citizens are not willing to pay the cost and change their way of life.

[3] There are exceptions, like the small group of people that travelled to Glasgow by bike and boat.

[4] But they are too few and, critics say, the bike and boat journey is actually not that much cleaner than flying.

[5] Ultimately, the people living in developed countries will have to make significant changes to their lifestyle, which includes, but is not limited to, transport.

[6] Meat consumption and fast fashion, to name but a few, are major contributors of emissions and limiting them requires a change in way of living.

[7] At the same time, several poor countries are entering economic and industrial growth and are expected to contribute increasingly to global emissions in the coming years.

[8] Proposals on how richer countries that caused the crisis are going to support the poor who suffer from it remain insufficient.

[9] Even the COP21 attendance reflects this with world leaders of large emitters declining and poorer developing countries not being able to afford sending large delegations to plead their case.

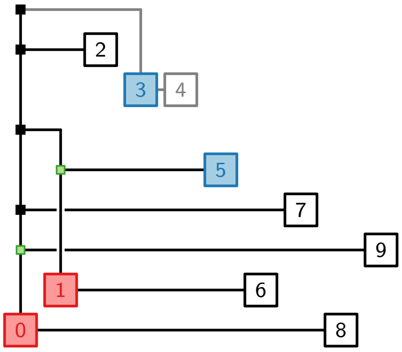

Figure 3: A news item and its story tree ST: the leaves of ST correspond to the sentences, ordered from left to right. The vertical axis indicates (from bottom to top) the sequence in which sentences become part of connected components of size at least two. Summary sentences chosen from salient leaves of ST are red, sentences chosen by k-center are blue; grey sentences are irrelevant side-stories which are pruned from the tree.

vertical edge in our drawing thus corresponds to the persistence of its primary pair. Note again that we are using purely the order of the merges and not the exact t values when drawing the tree.

A merge tree has an interior node for each merge; the node stores the corresponding t value. We say that an interior node is a **salient node** if its children are either two primary pairs or a primary pair and a secondary leaf. A primary pair in the sub-tree rooted at a salient node is a **salient primary pair**. At a salient node three or four sentences merge which are semantically closer to each other than to any other part of the text. We hence posit that these sentences describe a salient part of the input story. We further posit that a salient primary pairs represent central sentences of the story.

In the drawing we indicate salient nodes by a green square and all other interior nodes by a black square; their height corresponds to the ordered sequence of t values. Our example has two salient merge nodes and two salient primary pairs: $(0, 8)$ and $(1, 6)$. The primary pair $(3, 4)$ is not salient, we say it is an **irrelevant primary pair**. An irrelevant primary pair does not collect any further support in terms of secondary leaves before it merges into the component of a salient primary pair and dies. We posit that irrelevant primary pairs indicate irrelevant side stories and trim (remove) them from the story tree.

To conclude: A story tree is a merge tree with salient nodes and without irrelevant primary pairs. Section A in the appendix shows additional story tree visualizations.

4. Summarization via story trees

We posit that story trees succinctly represent the salient parts of a text. To showcase their potential, we introduce in Section 4.1 below three methods to construct extractive summaries via story trees.[1] In the remainder of this section we report on the setup and the results of an experimental study which evaluates the quality of the summaries constructed with our three methods both in an automated fashion against baselines and via human annotation.

4.1. Extractive summaries via story trees.

We posit that the number s of salient nodes captures the number of distinct story lines within a text. However, for long texts, s might be unreasonably large. We hence construct summaries of size $k = \min\{s, 5\}$.

Salient ST leaves [Salient STL] We choose one sentence from each salient primary pair. If there are more than 5 salient primary pairs, then we restrict ourselves to the 5 most persistent salient primary pairs. From each salient primary pair we choose the sentence that minimizes the average distance to the leaves of the story tree (all sentences minus the irrelevant pairs), following the intuition that central salient sentences are more representative.

k-center [k-center] We choose the k sentences that minimize the maximum distance to all other sentences in the input text.

k-center ST leaves [k-center STL] We choose the k sentences that minimize the maximum distance to all other sentences in the story tree. That is, we disregard irrelevant primary pairs.

[1] The code for visualization of story trees and the summarization methods will be made available at https://github.com/panteaHK/StoryTree.

4.2. Data

We use the CNN/Daily Mail data set (Hermann et al., 2015) to evaluate our extractive summaries. Each element in this data set consists of a news story and so-called “highlights”: manually curated summaries which partially consist of extracted sentences and partially of sentences composed by a human annotator. This CNN/Daily Mail data is partitioned into three parts: train, test, and validation. Our approach is based directly on the topological features of the input texts and is not learning-based. Hence we do not use the training part of the data set. The test data sets contains 11.490 elements and the validation data set contains 13.368.

To pre-process the new stories we use the `punkt` package from the `nltk` library (Bird, 2006) for sentence-tokenization and word-tokenization. Furthermore, we use the `nltk` library for stop-word removal. After preprocessing each sentence is now a bag-of-words. We use the pre-trained word2vec GoogleNews model[2] (Mikolov et al., 2013b; Mikolov et al., 2013a) to create sentence representations. Hence we remove all words which do not exist in the GoogleNews vocabulary. We then create the representation for each sentence by averaging the word2vec representations for all of its remaining words.

Very short sentences can be problematic for our persistence-based method: they often contain quite general words which make them similar to other sentences. As such they become part of spurious primary pairs which do not represent the start of a significant story line. For each news story we compute the average sentence length and the standard deviation σ. We are eliminating sentences which are shorter than $1.5\,\sigma$. Sentences which are (near to) duplicates of other sentences (figure captions or highlight boxes from the original news story) are equally problematic for our method. We are eliminating the second occurrence of any two sentences that have cosine distance less than 0.15. Figure 4 and Figure 5 show the story trees of two CNN/Daily Mail news stories that are pre-processed as described above. Further examples can be found in the appendix.

4.3. Evaluation

The CNN/Daily Mail data set contains highlights for each news story; we can compare our summaries from the story trees with these highlights. It is important to note that (i) our summaries are extractive summaries while the highlights are partially manually composed, and (ii) our summaries and the highlights do not necessarily have the same number of sentences. Nevertheless, we can measure the overlap between our summaries and the highlights using ROUGE scores (Lin and Hovy, 2003). However, since ROUGE scores are known to be heavily influenced by the exact wording of the target (the highlights), we also perform a manual evaluation which ranks our summaries with respect to each other and with respect to randomly created baseline summaries.

Automatic evaluation. We are using the elements (news story + highlights) of the test data set for our automated evaluation. We remove all elements with a news story of three or less sentences; this concerns a total of 11 elements. For the remaining 11.479 news stories we create summaries of k sentences with our three methods. Additionally, we create three random summaries of k sentences each by random sampling of the news story. We ensure that no two random summaries contain exactly the same k sentences. Note that a random summary might contain exactly the same sentences as one of our summaries.

Finally, we compute a so-called *upper baseline* summary for each news story, which is intended to approximate the best possible extractive summary of size k as measured by the ROUGE scores with the highlights as the target. We create the upper baseline summaries in a greedy manner by incrementally adding the sentence which increases the average of the three ROUGE scores the most.

We compute the F_1 scores for ROUGE-1, ROUGE-2 and ROUGE-L for all types of summaries. In Table 1 we report the respective scores averaged over all 11.479 news stories.

	R1	**R2**	**RL**
Upper baseline	53.2	31.1	36.4
Salient STL	31.6	10.5	19.5
k-center	29.8	9.1	18.3
k-center STL	29.4	8.9	18.1
Random	26.7	7.5	16.7

Table 1: The F_1 ROUGE scores in %.

Manual evaluation. We are using 30 random elements (news story + highlights) from the validation data set for our manual evaluation. The corresponding news stories (IDs 54 to 84, excluding 67) range from 13 to 57 sentences in length. As before we generate summaries of k sentences for each of our three methods as well as three random summaries of k sentences each. Three annotators carried out the manual evaluation. We use a pairwise ranking task to judge the summaries: annotators are presented with pairs of summaries and the source text; they have to indicate which summary is better. There are 6 different summaries; the annotators have to perform $\binom{6}{2} = 15$ pairwise comparisons per news story. If two or more summaries are identical, then we include only one copy in the evaluation, resulting in 10, or even just 6, pairwise comparisons.

To compute the pairwise **inter-annotator agreement** we use the Spearman correlation of their rankings in all comparison tasks. The average over all three pairwise agreement scores is **0.72**.

To evaluate the pairwise comparisons between different methods, we draw inspiration from ranked-choice voting, specifically, from the Schulze voting system (Schulze, 2011). For two summarization methods X and Y we define $d[X, Y]$ as the number of times a sum-

[2] `https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit`

[0] A self-driving car is gearing up for a 3,500-mile cross-country road trip from San Francisco to New York that begins next week.

[1] A person will always be behind the wheel of the autonomous car, developed by Delphi Automotive, in order to take control if there is a situation the vehicle cannot handle on its own.

[2] The car will begin its journey on March 22 so that it can reach New York in time for the International Auto Show.

[3] Delphi, a major Michigan-based auto supplier, is planning for the car to drive eight hours a day for eight days.

[4] The autonomous vehicle will mainly stick to the highways, where it will be able to pass slower cars and maintain a safe distance from other automobiles all on its own.

[5] Delphi CTO Jeff Owens told WIRED the road trip will be the 'ultimate test' for the vehicle, as it will be exposed to a 'range of driving conditions' as well as various types of weather.

[6] According to Delphi officials, the road trip will be the longest automated drive ever attempted in North America.

[7] In 2010 an autonomous van created by Italian company VisLab completed an 8,000-mile journey from Europe to Shanghai for three months.

[8] Unlike Google's driverless car (pictured), Dephi's vehicle senors are tucked in the car's front, rear and sides and thus cannot be seen.

[9] The Delphi vehicle has driven around the company's Silicon Valley office and has already completed a trip from San Francisco to Los Angeles.

[10] But the car really earned its stripes when, during a demonstration at the Las Vegas Consumer Electronics Show in January, it braked by itself as two drunk men stumbled into the street in front of it.

[11] Delphi says the vehicle is capable of making complex decisions, like stopping and then proceeding at a four-way stop, merging onto the highway or maneuvering around a bicyclist or a trash can.

[12] When the car wants the driver to resume control, it uses a verbal warning and flashes lights on the dashboard.

[13] Although most experts say a true driverless vehicle is at least a decade away, Delphi's autonomous automobile is helping the technology look more like a regular car.

[14] The car also has cameras throughout, including one that watches the driver.

[15] Because lidar systems can cost around $70,000 apiece, they drive up the price of the autonomous prototypes.

[16] But Delphi engineer Doug Welk said one of the reasons for the road trip is to help decide what combination of sensors is best suited for the car, which will ultimately help lower costs.

[17] Delphi estimates it will cost around $5,000 to make a vehicle almost fully autonomous by 2019.

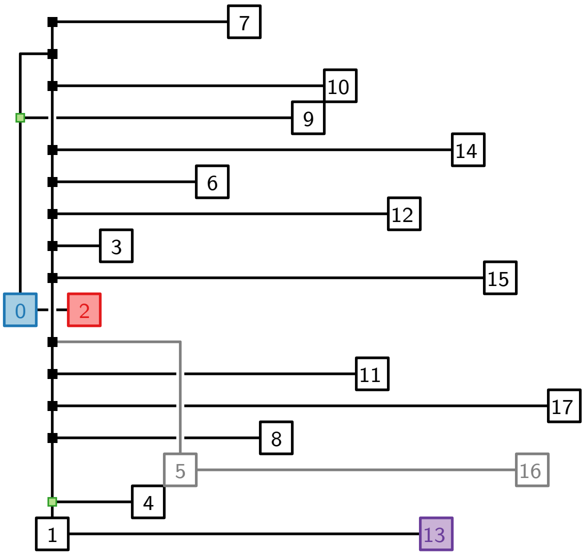

Figure 4: CNN/Daily Mail: *Validation set, ID 14*. Salient STL sentences are red, k-center sentences are blue, purple indicates sentences chosen by both; grey sentences are irrelevant side-stories which are pruned from the tree.

mary generated by X is preferred over a summary generated by Y over all news stories and annotators. If two summaries generated by different methods X and Y are the same, then we add 0.5 to both $d[X, Y]$ and $d[Y, X]$ for that comparison. Furthermore, we do not compare different summaries generated randomly; all randomly generated summaries (for a single news story) correspond to a single method (Random). The result is a matrix showing, for each pair of methods, how often one is preferred over the other. Since there are more comparisons in total with the Random method, we further normalize the values in the matrix to percentages as $\hat{d}[X, Y] = d[X, Y]/(d[X, Y] + d[Y, X]) \times 100\%$. Table 2 shows the resulting matrix of pairwise preferences. We can directly see that Salient STL is the preferred method, followed by k-center and k-center STL. All three of our methods clearly outperform Random.

The ROUGE score of the manually evaluated summaries are presented in Table 3. Note that the ROUGE scores of these summaries agree with the ordering of preferences of summaries calculated from the manual evaluations. This constitutes a validation of the automatic evaluation.

	Salient STL	**k-center**	**k-center STL**	**Random**
Salient STL	-	52	58	76
k-center	48	-	57	73
k-center STL	42	43	-	74
Random	24	27	26	-

Table 2: The pairwise preferences in %.

	R1	**R2**	**RL**
Salient STL	31.9	10.3	18.9
k-center	31.2	9.8	18.7
k-center STL	29.8	8.7	17.5
Random	27.9	8.4	17.5

Table 3: The F_1 ROUGE scores in %.

5. Discussion and Conclusion

In this paper we introduced story trees, a hierarchical structure that visually and semantically captures the salient shape of a story using persistent homology. We established story trees as a useful tool to create extractive summaries. In particular, our results show that summary sentences chosen based on salient story tree leaves outperform k-center extractions and random baselines when calculating ROUGE scores on the highlights of news stories from the CNN/Daily Mail data set. These

[0] Wellness guru Sarah Wilson has weighed in on the Belle Gibson controversy, saying too many health bloggers do not understand their power and responsibility.

[1] Ms Gibson, founder of the popular Whole Pantry app, has faced intense criticism since doubts were raised about whether she has terminal cancer.

[2] Ms Wilson, author of I Quit Sugar, told Daily Mail Australia that the incident was 'unfortunate' and that health bloggers have a huge responsibility to their followers.

[3] 'The real issue that's going on is there's a lot of people out there not taking on board the responsibility that comes with all this,' she said.

[4] Whole Pantry founder Belle Gibson last week said she would issue an explanatory statement, but it has not materialised .

[5] 'I don't know that you need to be a dietitian to share information about your health journey.

[6] 'If you are somebody who works online and you claim to be a (health) blogger you have to take on certain responsibilities.

[7] 'There's a lot of people out there who are't taking on the responsibility of being a disseminator of information.

[8] 'We've got to be very transparent and acknowledge the science is not categorical - but respect the science that is there.'

[9] Ms Wilson has said in a blog about the issue that 'diet can't cure disease'.

[10] Ms Gibson has stayed mum about her health since admitting in an interview with The Australian that her previous claims cancer had spread to her blood, spleen, uterus and liver may have been a 'misdiagnosis'.

[11] In the weeks since doubts were raised about Belle Gibson's cancer and charitable donations, her US and Australian publishers have withdrawn support for her recipe books .

[12] In the weeks since, her US and Australian publishers have withdrawn support for her recipe books.

[13] The Whole Pantry iPhone app has been removed from the Apple Store and her much-touted Apple Watch app is no longer showcased on the technology company's website.

[14] In an interview with Daily Mail Australia last Tuesday, Ms Gibson said she understood why many of her followers were upset, but she criticised the 'maliciousness' of some critics.

[15] She said she would release a statement addressing concerns about her, but that it was 'taking longer than anticipated' due to 'constantly arising issues' and articles she needed to stay on top of.

[16] Asked if the statement would address concerns about her medical history, she said 'Of course, and every which otherwise possible.'

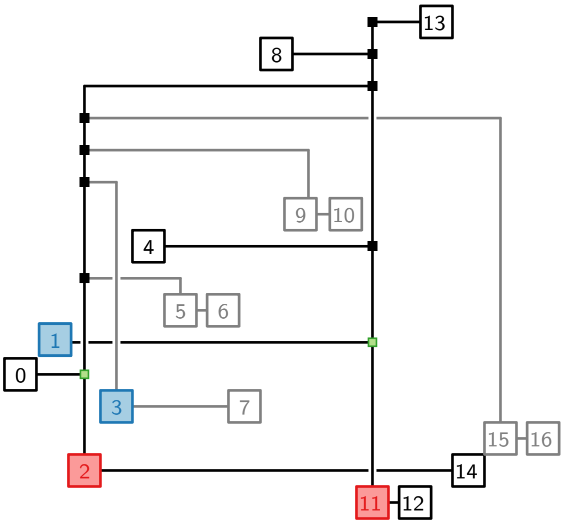

Figure 5: CNN/Daily Mail: *Validation set, ID 60*. Salient STL sentences are red, k-center sentences are blue, grey sentences are irrelevant side-stories which are pruned from the tree.

results were confirmed by a manual evaluation.

There are several possible improvements to our story trees. First of all, the tree structure currently depends solely on the order in which sentences connect; it does not take the persistence of a component or the time between two connections into account. This can affect the stability of the trees when one or more connections are formed essentially at the same time. Second, we currently define salient nodes as nodes whose children are either two primary pairs or one primary pair and a secondary leaf. These numbers appeared to work well with our news story corpus, however, longer text might benefit from different constants. We will explore ways to adaptively determine these values based on the length and possibly also the structure of the text.

Story trees can be created using any type of semantic representation. Our current approach represents each sentence by the average of the vector representations of its words. As such, it is very sensitive to name entities. Named entities tend to have a lower frequency in the training corpus and consequently the trained word embedding vector representing names have larger lengths (Schakel and Wilson, 2015). The larger vectors of named entities influence the sentence representation more than other words. Hence, sentences that include the same named entities tend to be very close to each other, regardless of the remainder of the sentences, and form primary pairs (see Figure 6). We plan to investigate various directions to mitigate this issue, for example, by using contextualized embeddings, or by using adaptive weights in the embeddings of names. During our experiments we encountered news stories for which neither of our methods could create a good summary. When inspecting these stories we found that we would not be able to create a summary by hand, for the simple reason that the texts in question were either incoherent collections of sentences (see Figures 7 and 8 in the appendix) or contained two or more disjoint story lines. We posit that story trees might be used as a tool to asses the quality of texts, we plan to explore this direction in the future.

Acknowledgments

This research is partially funded by the PhD in the Humanities Grant provided by the Netherlands Organization of Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO) PGW.17.041 awarded to Pia Sommerauer. We would like to thank anonymous reviewers for feedback that helped improve this paper. All remaining errors are our own.

6. Bibliographical References

Almgren, K., Kim, M., and Lee, J. (2017). Mining Social Media Data Using Topological Data Analysis.

In *Proceedings of the 2017 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 144–153.

Chiang, I.-J. (2007). Discover the semantic topology in high-dimensional data. *Expert Systems with Applications*, 33(1):256–262.

Christianson, N. H., Sizemore Blevins, A., and Bassett, D. S. (2020). Architecture and evolution of semantic networks in mathematics texts. *Proceedings of the Royal Society A*, 476(2239):20190741.

Doshi, P. and Zadrozny, W. (2018). Movie genre detection using topological data analysis. In *Proceedings of the 2018 International Conference on Statistical Language and Speech Processing (SLSP)*, volume 11171, pages 117–128.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Gholizadeh, S., Seyeditabari, A., and Zadrozny, W. (2018). Topological signature of 19th century novelists: Persistent homology in text mining. *Big Data and Cognitive Computing (BDCC)*, 2(4):33.

Guan, H., Tang, W., Krim, H., Keiser, J., Rindos, A., and Sazdanovic, R. (2016). A topological collapse for document summarization. In *Procceding of the 2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5.

Kushnareva, L., Cherniavskii, D., Mikhailov, V., Artemova, E., Barannikov, S., Bernstein, A., Piontkovskaya, I., Piontkovski, D., and Burnaev, E. (2021). Artificial Text Detection via Examining the Topology of Attention Maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Lee, Y., Barthel, S. D., Dlotko, P., Moosavi, S. M., Hess, K., and Smit, B. (2017). Quantifying similarity of pore-geometry in nanoporous materials. *Nature Communications*, 8(1):1–8.

Lin, C.-Y. and Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL)*, volume 1, pages 71—-78.

Mann, W. C. and Thompson, S. A. (1988). Towards a functional theory of text organization.

Marcu, D. (1999). Discourse trees are good indicators of importance in text. *Advances in automatic text summarization*, 293:123–136.

Michel, P., Ravichander, A., and Rijhwani, S. (2017). Does the Geometry of Word Embeddings Help Document Classification? A Case Study on Persistent Homology Based Representations. In *Proceedings of the 2nd Workshop on Representation Learning for NLP (RepL4NLP)*, page 235.

Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Munch, E. (2017). A user's guide to topological data analysis. *Journal of Learning Analytics*, 4(2):47–61.

Nakamura, T., Hiraoka, Y., Hirata, A., Escolar, E. G., and Nishiura, Y. (2015). Persistent Homology and Many-Body Atomic Structure for Medium-Range Order in the Glass. *Nanotechnology*, 26(30):304001.

Nielson, J., Paquette, J., Liu, A., Guandique, C. F., Tovar, C. A., Inoue, T., Irvine, K.-A., Gensel, J., Kloke, J., Petrossian, T., Lum, P., Carlsson, G. E., Manley, G., Young, W., Beattie, M., Bresnahan, J., and Ferguson, A. R. (2015). Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nature Communications*, 6:8581.

Reimers, N. and Gurevych, I. (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.

Savle, K., Zadrozny, W., and Lee, M. (2019). Topological Data Analysis for Discourse Semantics? In *Proceedings of the 13th International Conference on Computational Semantics - Student Papers*, pages 34–43.

Schakel, A. M. J. and Wilson, B. J. (2015). Measuring word significance using distributed representations of words. *Proceedings of the 5th International Conference on Artificial Intelligence and Security (ICAIS)*, abs/1508.02297.

Schulze, M. (2011). A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. *Social Choice and Welfare*, 36(2):267–303.

Zadrozny, W. W. (2021). A Note on Argumentative Topology: Circularity and Syllogisms as Unsolved Problems. *Computing Research Repository (CORR)*.

Zheng, H. and Lapata, M. (2019). Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247.

Zhu, X. (2013). Persistent homology: An Introduction and a New Text Representation for Natural Language Processing. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, page 1953–1959.

7. Language Resource References

Bird, S. (2006). NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Hermann, K. M., Kociskỳ, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching Machines to Read and Comprehend. In *Proceedings of the 28th International Conference* *on Neural Information Processing Systems (NIPS)*, volume 1, pages 1693–1701.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceeding 1st International Conference on Learning Representations, (ICLR) 2013*, pages 2–4.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American chapter of the Association for Computational Linguistics (NAACL) : Human language technologies*, pages 746–751.

A. Story Tree Examples

[0] England fans should hide behind the sofa rather than watch the Ashes this summer, says Mark Butcher.

[1] After a humiliating World Cup exit, England face a Test series in the West Indies, followed by series at home to New Zealand and Australia.

[2] Speaking to Radio 5 Live, the former England opener warned fans to expect an 'absolute hiding' and a 'horrendous six months'.

[3] Butcher, who played in 71 Tests for England and scored eight centuries, added: 'When Australia come, just don't watch, hide behind the sofa.' England's under siege captain and coach Eoin Morgan (left) and Peter Moores speak at the SCG.

[4] Former England opener Mark Butcher has predicted the worst for the national team in the near future.

[5] Here England are preparing for a dead rubber against Afghanistan, with Peter Moores' men having far more to lose than just a meaningless group match.

[6] The indications are that Moores will keep his job, at least for the Test series in the Caribbean, whatever happens at the Sydney Cricket Ground tomorrow, but the last thing he needs is another desperate defeat by minnows.

[7] The fact England turned up early yesterday for what used to be known as naughty-boy nets and had a full training session said everything about their need to take something from their worst-ever World Cup.

[8] Victory against Afghanistan would perhaps give Moores breathing space before the squad is announced on Tuesday for the three-Test series in the West Indies next month.

[9] England will be without Moeen Ali and Chris Woakes, two players who can still expect a one-day future, and it is far from certain that both will be able to take their places on the plane to St Kitts on April 2.

[10] Moores attempts to rally his troops on the SCG on Wednesday ahead of their match against Afghanistan .

[11] Moeen strained an abdominal muscle when bowling in the defeat by Bangladesh while Woakes has a 'stress reaction' in his left foot.

[12] It is to be hoped England give games to the two players yet to appear in this World Cup: James Tredwell and Ravi Bopara, who could then be making their final ODI appearances.

[13] After the World Cup England must start from scratch with a one-day squad who will be at or near their peak in 2019.

[14] That means there is little point in persevering in 50-over cricket beyond tomorrow with the likes of Tredwell, Bopara, Ian Bell, Stuart Broad or Jimmy Anderson, who said on Wednesday he wants to play on.

[15] His is not a view shared by former England wicketkeeper Matt Prior.

[16] Young England quick Chris Jordan works up a sweat during a tough day of training in the Sydney sunshine .

[17] In his Evening Standard column Prior wrote: 'Men like Ian Bell, Jimmy Anderson and Stuart Broad... remain key members of the Test side, so perhaps it is time for them to concentrate solely on that form of the game, especially with an Ashes summer approaching.' There is a familiar figure at the helm of the Afghan team in coach Andy Moles, the former Warwickshire batsman.

[18] He would have no qualms about piling on the misery for his homeland and said: 'We still believe we've got a scare in us in this World Cup, especially if we show composure at the top of the order.

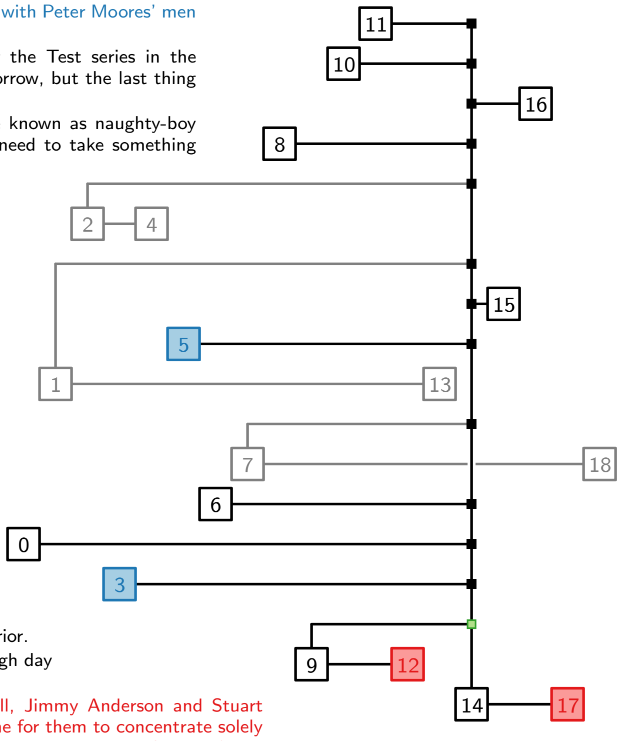

Figure 6: A sports news item and its story tree: Sentences are close due to named entities. The salient pairs $(12, 9)$ and $(14, 17)$ are due to the names of cricketers and the word "England"; the pair $(1, 13)$ is due to the term "World Cup" and country names.

[0] David Cameron is seen as 'arrogant' while Ed Miliband is 'weak' according to exclusive new polling for MailOnline which asks voters to sum up political leaders in a single word.
[1] For the first time voters have been captured on video making clear what they really think of the people bidding to run the country.
[2] Using webcam technology and surveying more than 2,000 people, the poll also reveals that Lib Dem leader Nick Clegg is seen as 'weak' while Ukip's Nigel Farage is branded 'racist'.
[3] Exclusive video polling for MailOnline reveals what voters think of the party leaders in their own words .
[4] The political blink test is designed to test the instant reaction voters have to seeing the party leaders jostling for their support on May 7.
[5] More than 2,000 people were surveyed by Populus for MailOnline, and each were given the option to record their responses on camera.
[6] It gives a Gogglebox-style insight into the personal reactions of members of the public to senior politicians.
[7] The most frequent response to Mr Cameron was 'arrogant', followed by 'smug', 'leader', 'liar' and 'Prime Minister'.
[8] Three people branded the Prime Minister a 'plonker', two called him 'evil' and one replied 'Blair'.
[9] Nine called him a w*nker, 11 said 'knob' and 8 used the words 'arsehole'
[10] Six mentioned Eton and four called the Prime Minister a 'c***'.
[11] When a photograph of Mr Miliband was displayed, the most popular response was 'weak' followed by 'idiot', 'Labour', 'useless' and 'untrustworthy'.
[12] Other frequent responses were 'weird', 'incompetent', 'boring' and 'honest'.
[13] The top 10 is completed by 'Wallace', the plasticine character the Labour leader is often likened to.
[14] Five people thought of 'bacon sandwiches' as the first reaction to Mr Miliband, after his disastrous attempt to be photographed eating breakfast a year ago.
[15] Five also called him a 'clever', four mentioned 'hope' and one said Mr Miliband was 'cute'.
[16] One said 'anus', two said 'crap' and five called him a 'moron'.
[17] The Lib Dem leader has seen his party's popularity collapse since entering coalition five years ago, in large part as a result of the broken promise to abolish tuition fees.
[18] Popular responses to the Deputy PM also include 'liar', 'don't know', 'untrustworthy' and 'useless'.
[19] In a sign that not everyone knows who he is, or is unsure what to feel about him, the top 10 is completed by 'idiot', 'nothing', 'who?'
[20] Two people described the Deputy PM as a 'poodle', four said he was a 'sell-out' and one branded him a 'tosser'.
[21] One called him a 'slag', another said 'numbskull' and two said 'pillock' Popular responses to Lib Dem leader Nick Clegg included 'liar', 'don't know', 'untrustworthy' and 'useless' The overwhelming reaction to Mr Farage was to brand him a 'racist'.
[22] Ukip has been dogged by revelations about its candidates and members, with Mr Farage insisting he will take action against those who embarrass the party.
[23] But he has also courted controversy by defending a candidate taped mocking 'poofters', referring to a Chinese woman as a 'Chinky bird' and threatening to 'shoot peasants'.
[24] The next most frequent reactions to Mr Farage included 'idiot', 'dangerous', 'arrogant' and 'bigot'.
[25] The top 10 responses also included 'untrustworthy', 'honest' and 'joke'.
[26] Five people said 'frog', five said 'yuk!'
[27] The overwhelming reaction to Ukip leader Nigel Farage was to brand him a 'racist'.
[28] Others called him an 'idiot', 'dangerous', 'arrogant' and 'bigot' Daisy Powell-Chandler, a consultant at Populus, said: 'It is a fact ignored by most politicians that the public aren't paying attention most of the time.
[29] 'Voters have families, jobs and bills that are much more interesting to them than today's policy launch or staged visit.'
[30] As the election approaches on May 7, more information will start trickle through.
[31] 'But many of the electorate will be choosing who to vote for (or whether to vote at all) based on how each party leader scores on the 'blink test'– the immediate reaction when they see or hear party leaders in the media.
[32] 'It is clear from the results that all political leaders are polarising figures with strong downsides.
[33] People think Cameron is arrogant, Miliband is weak, Clegg cannot be trusted, and Farage is a racist.
[34] 'But the flipside of these criticisms are things we like about them: Cameron may come across as smug but he is smooth and competent– an assertive leader.
[35] Clegg is the likeable liberal and Farage a charismatic, funny man.
[36] 'Over the course of the campaign, the electorate need to decide which of these attributes are most important and the emotional connections that the public make with these leaders are likely to be an important factor in deciding who the next Prime Minister is.'

Figure 7: Text of CNN/Daily Mail: *Validation set, ID 59*. This news story is a collection of irrelevant sentences; it does not have a story line and hence also no suitable summary. We show its story tree in Figure 8.

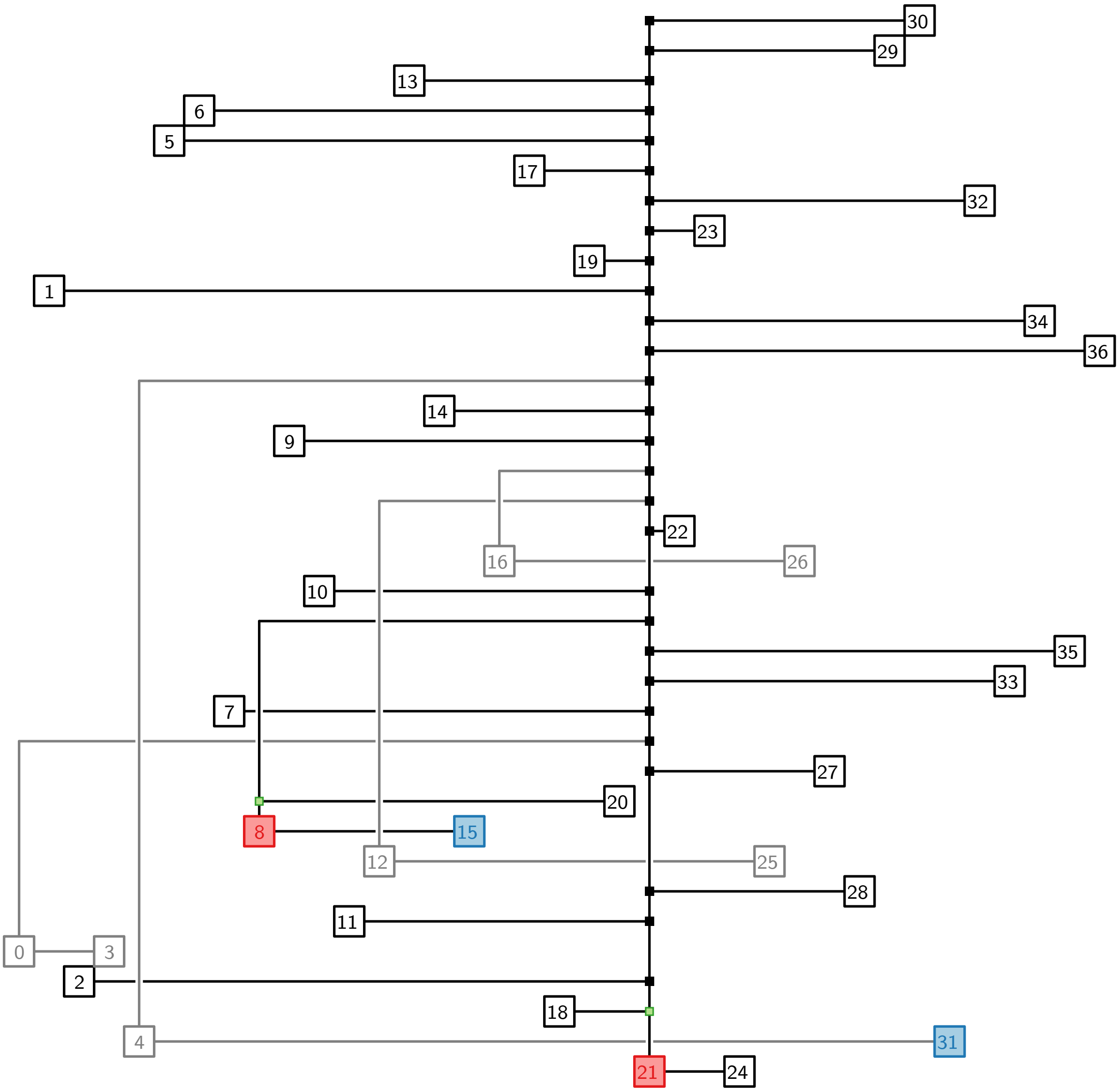

Figure 8: Story tree of CNN/Daily Mail: *Validation set, ID 59*, text see Figure 7.

[0] A man convicted of killing the father and sister of his former girlfriend in a fiery attack on the family's Southern California home was sentenced to death on Tuesday.

[1] Iftekhar Murtaza, 30, was sentenced for the murders of Jay Dhanak, 56, and his daughter Karishma, 20, in May 2007, the Orange County district attorney's office said.

[2] Murtaza was convicted in December 2013 of killing the pair in an attempt to reunite with his then-18-year-old ex-girlfriend Shayona Dhanak.

[3] She had ended their relationship citing her Hindu family's opposition to her dating a Muslim.

[4] To be executed: Iftekhar Murtaza, 30, was sentenced to death Tuesday for the May 21, 2007 murders of his ex-girlfriend's father and sister and the attempted murder of her mother .

[5] Authorities said Murtaza and a friend torched the family's Anaheim Hills home and kidnapped and killed Dhanak's father and sister, leaving their stabbed bodies burning in a park 2 miles from Dhanak's dorm room at the University of California, Irvine.

[6] She was stabbed and left unconscious on a neighbor's lawn.

[7] Murtaza was interviewed by police several days later and arrested at a Phoenix airport with a ticket to his native Bangladesh and more than $11,000 in cash.

[8] Jurors recommended that Murtaza be sentenced to death for the crimes.

[9] Attack: Murtaza torched his ex-girlfriend's family's Orange County home after they broke-up, believing the murders of her family would reunited them .

[10] Religious differences: Murtaza dated Shayona Dhanak when she was 18 in 2007.

[11] She broke up with him when her Hindu parents allegedly told her they would stop paying her college tuition if she continued to date the Muslim man .

[12] Two of his friends were also sentenced to life in prison for the murders, but one of them, Vitaliy Krasnoperov, recently had his conviction overturned on appeal.

[13] Authorities said Krasnoperov hatched the plot to kill the Dhanaks with Murtaza and tried to help him hire a hit man.

[14] They said another friend, Charles Murphy Jr., helped Murtaza carry out the killings after Dhanak said she planned to go on a date with someone else.

[15] During the trial, Murtaza testified that he told many people he wanted to kill the Dhanaks because he was distraught over the breakup, but he said he didn't mean it literally.

[16] Didn't do it alone: Two of Murtaza's friends have been convicted in connection to the killings.

[17] Killer: Leela Dhanak testified how Iftekhar Murtaza, seen in this August photo, murdered her husband and elder daughter in a failed attempt to win over her younger daughter.

[18] Bloodbath: Autopsy reports showed Jayprakash Dhanak (left) suffered 29 stab wounds to his body, while a pathologist testified that Karishma Dhanak (right) was alive when her throat was slit and her body set alight .

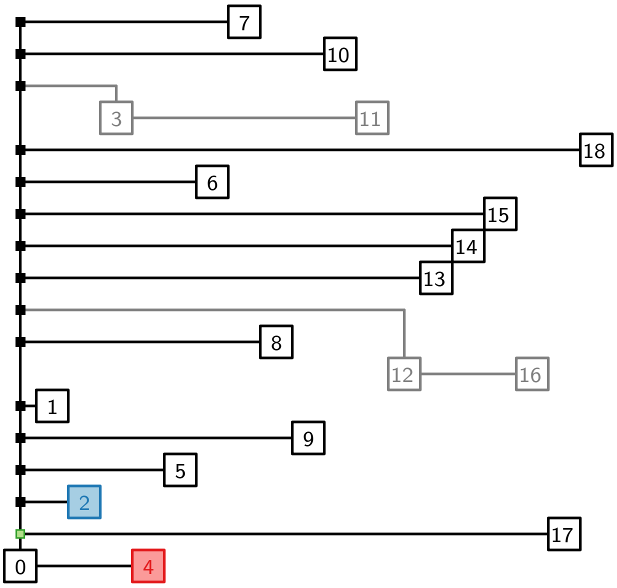

Figure 9: CNN/Daily Mail: *Validation set, ID 2.*

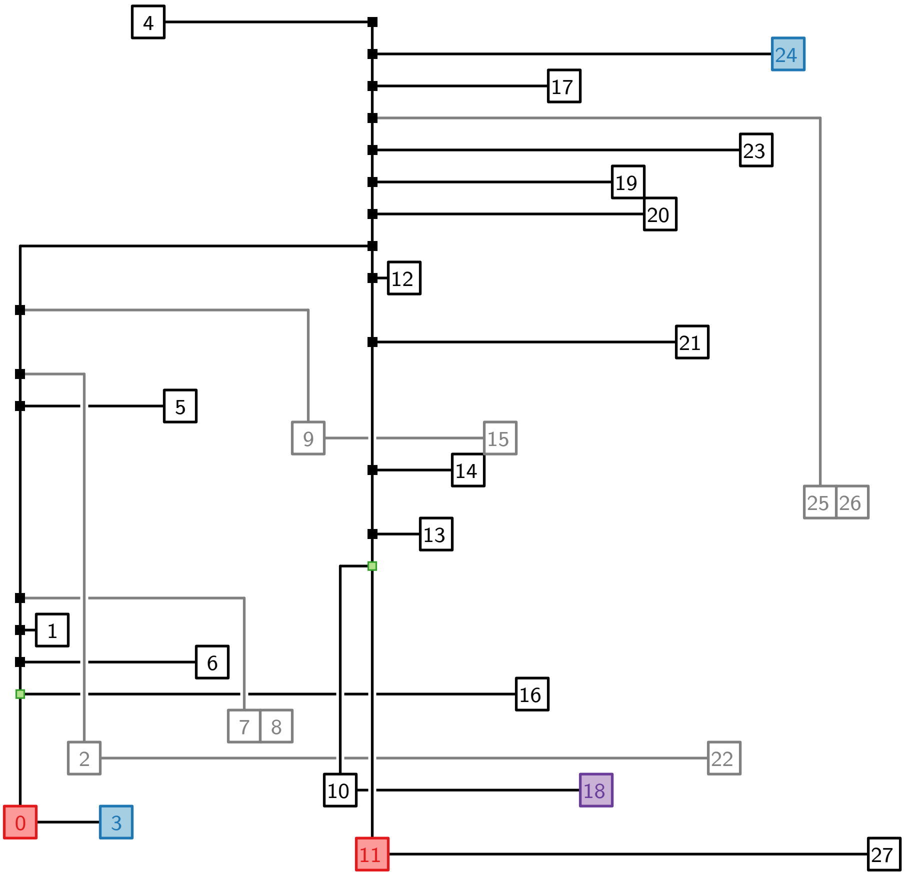

Figure 10: CNN/Daily Mail: *Validation set, ID 16*.

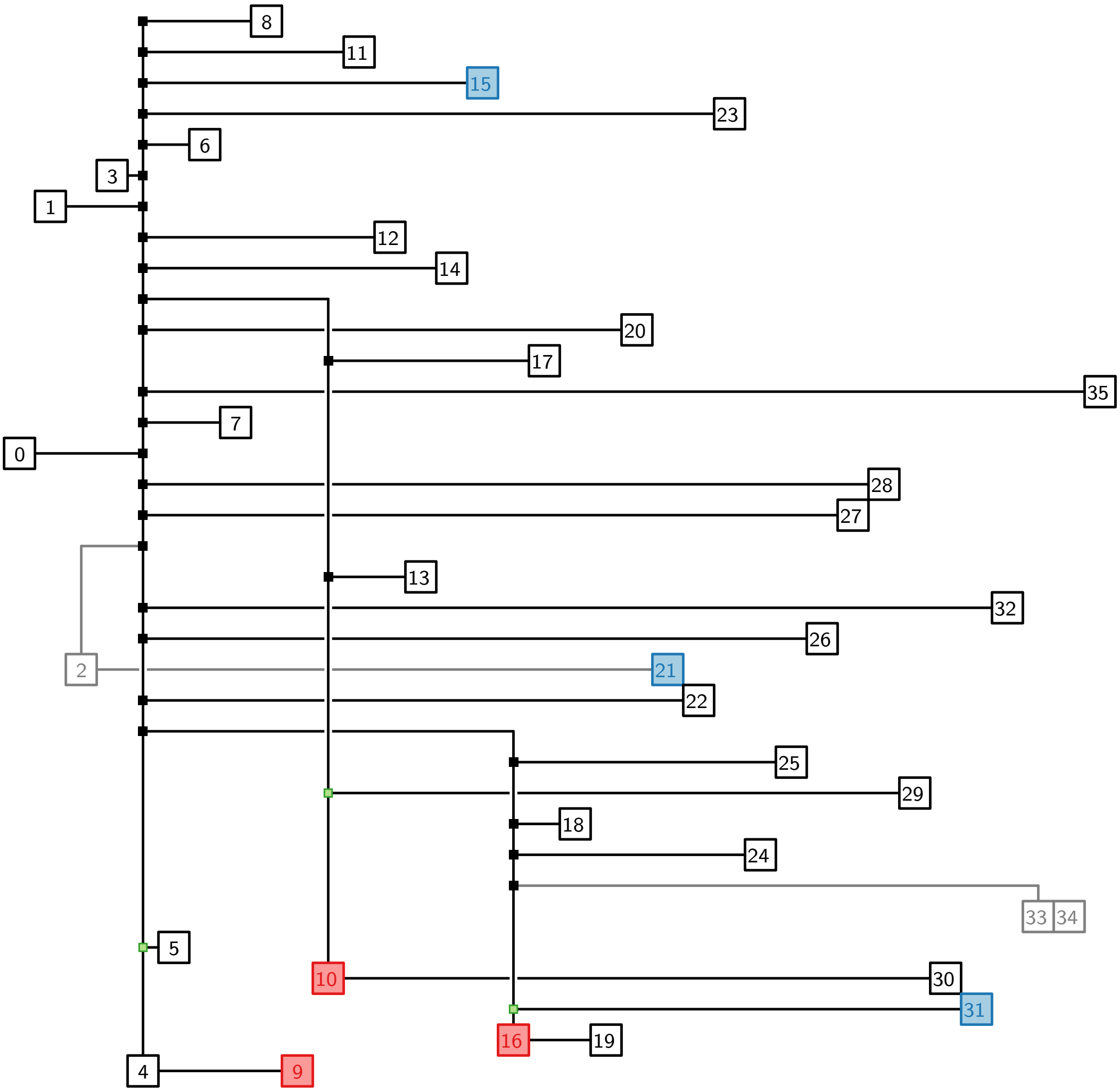

Figure 11: CNN/Daily Mail: *Validation set, ID 6*.

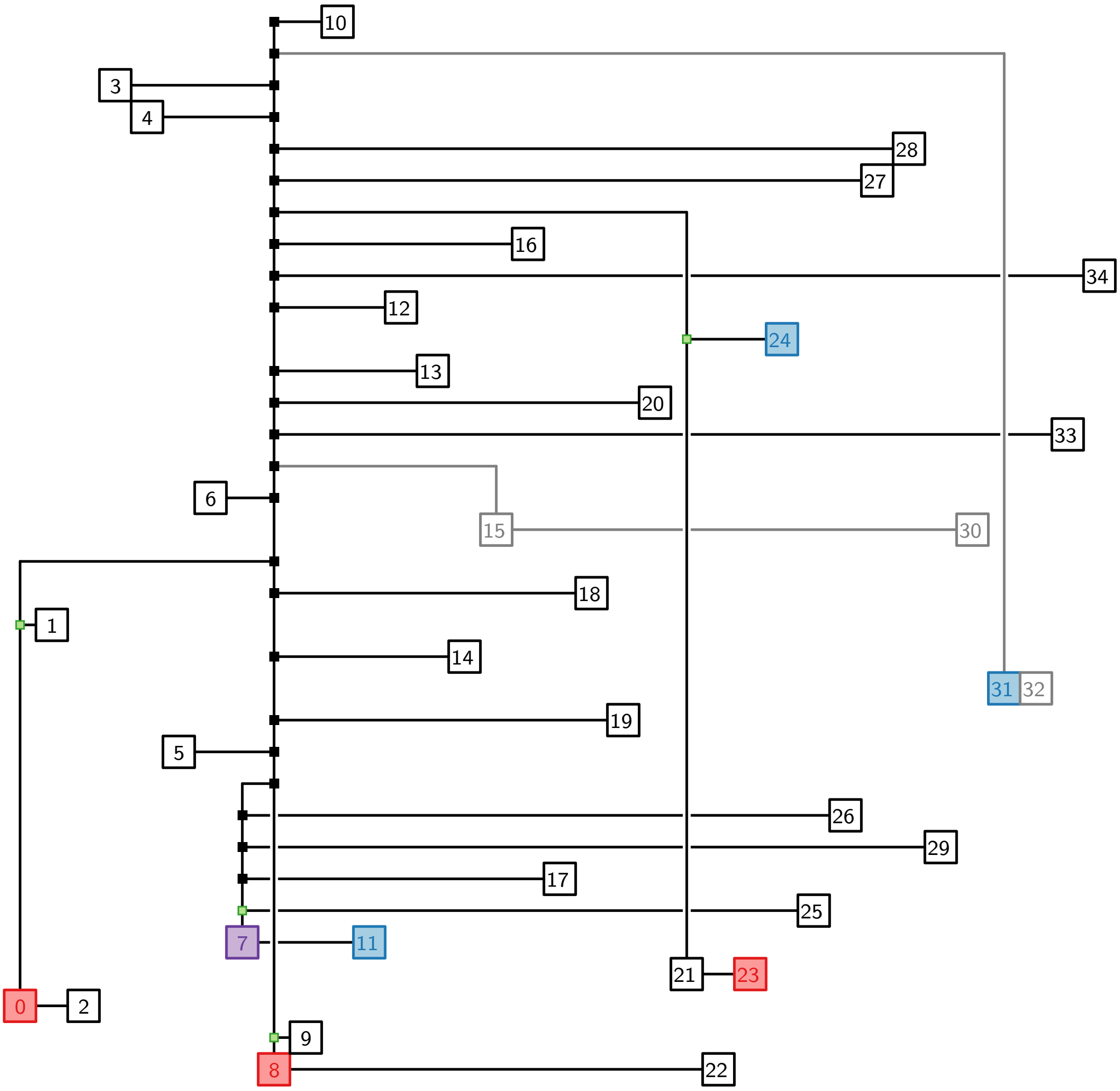

Figure 12: CNN/Daily Mail: *Validation set, ID 76*