CANarEx: Contextually Aware Narrative Extraction for Semantically Rich Text-as-data Applications

Nandini Anantharama[1]**, Simon D. Angus**[1,2] **and Lachlan O'Neill**[3]

[1]SoDa Laboratories, [2]Dept. of Economics, [3]Faculty of IT, Monash University, Australia

{nandini.anantharama, simon.angus, lachlan.oneill}@monash.edu

Abstract

Narrative modelling is an area of active research, motivated by the acknowledgement of narratives as drivers of societal decision making. These research efforts conceptualize narratives as connected entity chains, and modeling typically focuses on the identification of entities and their connections within a text. An emerging approach to narrative modelling is the use of semantic role labeling (SRL) to extract Entity-Verb-Entity (E-V-Es) tuples from a text, followed by dimensionality reduction to reduce the space of entities and connections separately. This process penalises the semantic richness of narratives and discards much contextual information along the way. Here, we propose an alternate narrative extraction approach - CANarEx, incorporating a pipeline of common contextual constructs through co-reference resolution, micro-narrative generation and clustering of these narratives through sentence embeddings. We evaluate our approach through testing the recovery of "narrative time-series clusters", mimicking a desirable text-as-data task. The evaluation framework leverages synthetic data generated using a GPT-3 model. The GPT-3 model is trained to generate similar sentences using a large dataset of news articles. The synthetic data maps to three topics in the news dataset. We then generate narrative time-series document cluster representations by mapping the synthetic data to three distinct signals synthetically injected into the testing corpus. Evaluation results demonstrate the superior ability of CANarEx to recover narrative time-series through reduced MSE and improved precision/recall relative to existing methods. The validity is further reinforced through ablation studies and qualitative analysis[1].

1 Introduction

Narratives are accounts of connected events, and are a basic mechanism of communication and of understanding our world (Piper et al., 2021). "Narrative modelling" is the discovery of underlying narratives (such as "politician is liar" or "inflation is high"), as well as the analysis of how their shape and prominence shifts over time (Zhang et al., 2019; Ash et al., 2021). Narrative modelling offers the potential to go beyond simple issue tagging or topic modelling to get at the richer question of how ideas are connected and framed, and when these relationships form, arise, and change.

In the realm of the social sciences, quantitative and causal analysis relies on structured, tabular data, with researchers recognising the potential for 'text as data' to greatly expand the realm of analysis into news, opinion, and political discourse (Gentzkow et al., 2019). In particular, if semantically related narratives can be reliably identified in a large corpus over time using narrative modelling, then established time-series techniques in the social sciences can be leveraged in a wide array of applications (Box-Steffensmeier et al., 2014). Recent work on narrative extraction such as Relatio (Ash et al., 2021) demonstrates the potential of such approaches to discourse analysis, when based on more traditional NLP techniques. The advances in Transformer-based language models such as BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) present the opportunity to vastly increase the power of such methods, incorporating a greater understanding of "context" into the model. Context is especially important to narratives since they are fundamentally composed of sequences of *coupled* language components.

We demonstrate and evaluate, CANarEx, a contextually-aware narrative modelling approach which employs state-of-the-art Transformer models to unlock the potential of the context-sensitivity of these models. We show the superiority of CANarEx's contextually-aware narrative modeling technique to (already powerful) existing methods in the literature. We discuss how these models can

---

[1]Code: https://github.com/nandinisa/CANarEx

be used to conduct previously infeasible analysis of the shifts in narratives through an illustration in the social sciences by understanding social discourse over time. We employ generative text models tuned on narratives extracted from original news, opinion and Parliamentary speeches to develop a highly realistic narrative time-series recovery evaluation task.

2 Related Work

Automated extraction of narratives and other artifacts for text as data is an area of ongoing research (Piper et al., 2021), spanning from classical NLP techniques to the new transformer architectures. Given the typical multi-entity, discursive nature of most narratives, identifying entities and interactions between entities are common building blocks of narrative modelling approaches. These interactions can be framed as semantic triples extracted using a semantic parser and can take forms such as {Subject-Predicate-Object}, {Entity-Verb-Entity}, with the key idea being that the subject and object of the triple represent entities and the verb or predicate of triplet represents the interaction (or event) between them (Rospocher et al., 2016; Spiliopoulou et al., 2017; Ash et al., 2021). These E-V-Es provide a succinct summary of the underlying events, actors addressing "who does what to whom" (He et al., 2015) and thus provide a framework that enables further downstream tasks. In terms of the "minimal model of narrativity" (Piper et al., 2021), the triples can be mapped to the features 'agents' and 'events'.

Zhang et al. (2019) propose a narrative modelling framework with a particular focus on identifying economic activities (debates, conflicts, compromise) around social processes such as industrial regeneration. Towards this end, their work prioritises identifying the points of view of each entity called attribution (beliefs, thoughts, speeches) in a given narrative alongside identifying the entities and events involved. Their pipeline includes manually annotated custom event labelling, pre-trained entity recognition(NER) (Dernoncourt et al., 2017) models and a semantic role labelling (SRL) ensemble comprising of Semafor (Das et al., 2014) and DeepSRL (He et al., 2017) for event and attribute extractions of the narratives. The results from the semantic frames (predicate with entities) is split into entities, extracted with the NER, and events through mapping to the custom event labels which

has the event type attribution. Through this process, their framework is able to attribute the statements and intents within a narrative to specific actors.

The 'Relatio' framework takes a more generalized approach to narrative modelling, seeking to extract narratives from any text data, and without extensive custom labelling. 'Relatio' (Ash et al., 2021) framework enables identifying the opinions, inflection points and changing trends from public discourse through narrative extraction from news. They use the {Entity-Verb-Entity} narrative structure, or the E-V-Es using SRL (Stanovsky et al., 2018) to identify the entities in the discourse. They have an additional step to convert the extracted high dimensional entities to lower dimensional entities using clustering of word vectors (Mikolov et al., 2013; Cer et al., 2018). The workflow is presented in Figure 1. The separation of the entities from their verbs whilst clustering results in loss of contextual information.

As seen from these examples, modelling the multi-entity, discursive nature of narratives typically involves identifying entities and interactions between these entities, with the rest of the pipeline reflecting the specific goals of the model. The implementations of these different tasks of the pipeline have evolved to reflect the new transformer architectures (Vaswani et al., 2017) and associated improvements. Wankmuller (Wankmüller, 2021) provided a review of the possibilities of transfer learning with transformers in the domain of social science studies, highlighting the higher prediction power and increased efficiencies of these pretrained language models (PLMs) compared to the historically common supervised learning models. The CANarEx framework builds on the work of Relatio by incorporating more features to support complex narratives, and by updating the existing components to leverage transformer based models.

3 CANarEx

Here we propose the 'Contextual narrative extraction' (CANarEx) framework that enables contextual narrative extraction. The framework is represented in Figure 3.

Contextual approach

CANarEx provides several improvements to the original Relatio framework. The entities of a narrative present themselves through multiple sentences and thus have prior references to itself and to other
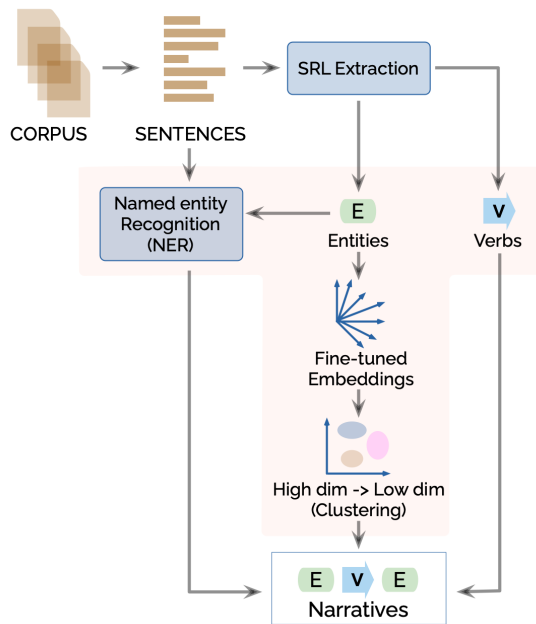
Figure 1: Adapted from Figure 3: Program Flowchart: Relatio Narrative Mining System, Ash et al., 2021 (Ash et al., 2021)

entities. These prior references need to be resolved in order to recover the context for the specific narrative. The problem of prior references is solved through the addition of co-reference resolution to the pipeline to translate all the subsequent mentions in a document to their respective named entities, pronouns and noun phrases. We use the SpanBERT-base model from Joshi et al. (Joshi et al., 2019b,a) with domain set to newswire ('nw') for this step.

Next, the "interactions" between these entities have to be recovered in order to formalize the actions being performed between entities. Relatio views this as a predicate-argument resolution problem, and leverages SRL to extract this structure, with the target outcome of identifying E-V-Es embedded in the document. For SRL, we use the AllenNLP framework (Gardner et al., 2017) specifically the BERT-based models introduced by Shi and Lin (2019). The SRL results in a frame file of predicate constructs with the numbered arguments (*ARGx* for example) representing entities interacting via predicate V (verb). This translates to how the entities are interacting, when and where. In CANarEx, the prior application of co-reference resolution ensures a more comprehensive E-V-E structure in this SRL step compared to the vanilla Relatio approach.

Micro-narratives

In the next step, we further split the extracted E-V-Es into micro-narratives. Micro-narratives are narratives nested within other narratives, and are referred to as narrative levels in the minimal model of narrativity (Piper et al., 2021). SRL can generate multiple frames per sentence, one for each connecting verb. These frames can contain nested or overlapping narratives. We identify and disambiguate these constituent micro-narratives by checking if the extracted E-V-E from a frame is a subset of another E-V-E from another frame of the same sentence. If yes, we split the superset narrative into its constituent micro-narratives, as demonstrated via an example in Figure 2. These generated micro-narratives are cleaned using stopwords removal.
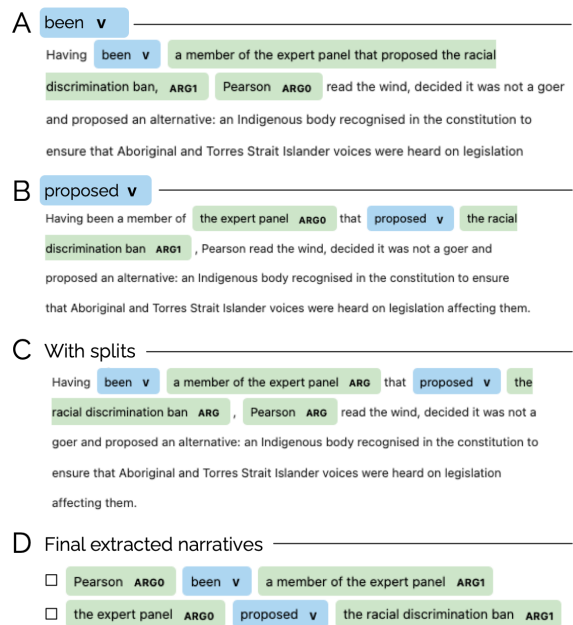

Figure 2: Micro narratives extraction. Two hinge verbs are identified in the sentence, been (A) and proposed (B). By splitting, both verbs are identified (C) and co-referencing resolves correctly the entities to provide the final narratives (D).

Narrative clustering

Next, we need to cluster the extracted artifacts to generate semantically similar groupings of narratives. Prior work in Relatio clustered the entities to achieve this goal. However, this fragmented approach ignores the connections between the entities and therefore loses the narrative aspect represented by the verb. We propose a modified approach that preserves contextual information - we use sentence embeddings using SBERT (Reimers

and Gurevych, 2019; Song et al., 2020) of the generated micro-narratives (E-V-Es), and perform clustering on these embeddings.

Optional - filtering of micro-narratives

The use of co-reference resolution and micro-narratives generation results in a very large number of E-V-Es relative to the original text corpus. A proportion of these might be repetitive, necessitating a subsequent step in reducing this collection into a viable subset without loss of information. To filter out such E-V-Es, we evaluate Document level clustering (TopN) and Textrank algorithms as two filtering approaches.

The TopN document clustering approach assumes that each document has a main theme and a sub-theme. Therefore, we cluster all the sentences within each document into two clusters, pick the largest of the clusters to be the main theme, and then pick top N sentences closest to the centroid of this main theme cluster of the document.

Another approach is using the Textrank algorithm which is an automated graph based summarisation technique (Mihalcea and Tarau, 2004). This approach involves representing text as a graph with nodes (sentences) and edges (similarity between the connecting sentences). The top sentences are determined using a form of paging algorithm (Page et al., 1999), i.e. the most connected sentences. The sentence similarity can be calculated using common substrings, cosine similarity with TF-IDFs or sentence embedding (Barrios et al., 2016; Kazemi et al., 2020). We use the version by Barrios et al. (2016) with a modified Okapi-BM25 (information retrieval ranking) function, and vary the ratio of summary in proportion to the co-referenced sentences within a document as 0.05, 0.1, 0.2 and 0.4.

Both these approaches yield the top sentences within the documents which are then used to extract micro-narratives. We perform a final clustering step on the filtered set of micro-narratives and choose the cluster with the largest silhouette score.

Data

We demonstrate the CANarEx framework by applying it on two major sources of data - Hansard[2] and Dow Jones Factiva[3] datasets. The goal was to
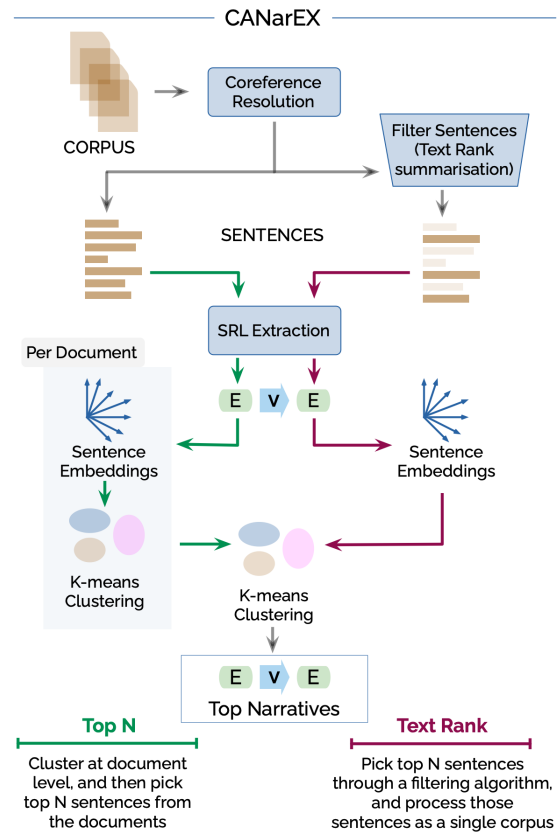
Figure 3: The CANarEx framework (including filtering for top narratives)

understand narratives of disadvantages in Australia, focusing on the group, 'First Nations' (Aboriginal and Torres Strait Islander people). The choice of these 2 data sources reflects two perspectives of the discourse - the debates by policymakers in the parliament (Hansard) and the reporting and editorials on the same topics by the news media (Factiva).

For the Hansard dataset, we filtered for the 'First Nations' group in speeches based on the presence of keywords: "first peopl", "first nations", "traditional own", "indigen", 'aborigin", and "torres strait island". We have made this dataset available as supplementary material.

We had 10554 (461180 sentences) of articles from Factiva and 7781 documents (440169 sentences) from Hansard describing the 'First Nations' group.

The results of the CANarEx framework are presented per dataset in Table 1. Each row of the tables represents the progression through each step of the CANarEx framework. A vast majority of sentences in both datasets (69% Factiva and 79% Hansard) were changed by co-reference resolution, underlining the importance of retaining context and
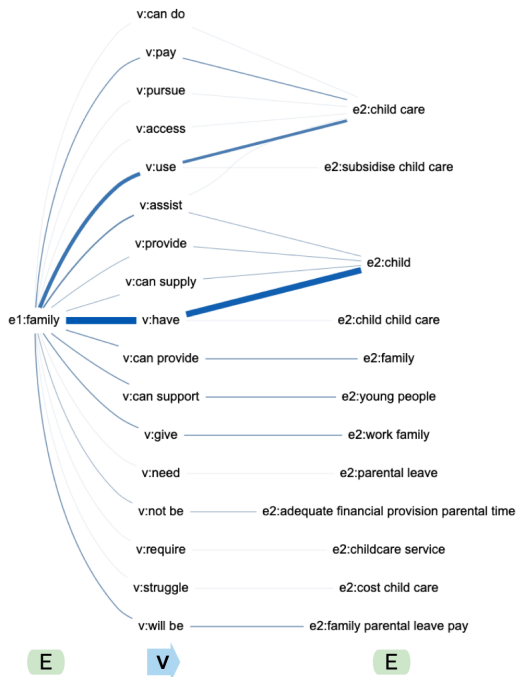
Figure 4: CANarEx generated micro-narratives on topic of young families, derived from the Hansard dataset.

	Factiva	Hansard
Total # sentences	461180	440169
After co-ref	455962	433117
No change	142228	92322
Changed	313734	340795
After SRL		
With a frame	418101	423784
Without frame	37861	9333
Micro-narratives		
Extracted Micro-Narratives	389429	404237
Cleaned narratives	294418	290360
Optional filtering		
With textrank	15221	82294
With topN	10516	7608

Table 1: Factiva and Hansard results

entities through this step. The application of SRL resulted in frames being generated in over 90% of these sentences. Using these SRL extracted frames, we were able to generate 389429 micro-narratives from the Factiva dataset and 404237 from Hansard (retaining only the frames with a minimum of two entities). These were further cleaned using stopwords removal and removal of the verbs {*is, are*}, reducing the micro-narratives count by about 25%. The optional filtering step was applied on these cleaned sets of micro-narratives to arrive at the result set.

The result of Textrank filtering on the topic of young family within the 'First nations' Hansard data is shown in Figure 4. The links between the E-V-Es are weighted by frequency of occurrence. This enables us to identify the dominant micronarratives per entity pair, such as the higher occurrence of *e1:family - v:use - e2:child care* compared to other micro-narratives. The representation is inspired by the transformer attention model mechanism (Vaswani et al., 2017). A Factiva example is provided in Appendix C.

4 Evaluation

We evaluate the performance of CANarEx by testing the recovery of the time-series narrative across the corpus. To do this, we rerun our pipeline on

time-series clusters of synthetic data and compare the results with the output of Relatio, our baseline model.

GPT-3 fine-tuning

GPT-3 is fine-tuned to generate paraphrased sentences using similar sentence pairs as training data. We use the original sentences extracted from the Factiva documents, and filter for sentences that remain similar post co-reference resolution. We do this to ensure that the samples generated from GPT-3 contain sentences with references to pronouns and entity mentions. We then perform SBERT's (Reimers and Gurevych, 2020) paraphrase mining on these filtered sentence embeddings. This results in semantically similar sentence pairs [x,y], the training data for GPT-3. We also set a threshold of $0.75 \leq$ cosine score < 0.99 and having word length > 5 resulting in 4961 [x,y] sentence pairs. We fine-tune the Ada GPT-3 model over these sentence pairs over 4 epochs. This is presented in Figure 5 (a).

Once the GPT-3 model is fine-tuned, we use this to generate synthetic test sentences.

Synthetic test data

The test sentences contain both topic cluster sentences and noise sentences. For topic sentences, we sample sentences from 3 chosen topic clusters from the Factiva dataset. These 3 clusters represent the topics: climate change, Kevin Rudd and indigenous people, Figure 6 (a). Noise sentences were sampled from clusters other than the chosen topic clusters. Topic cluster sentences were sampled from the closest euclidean distance to centroid,
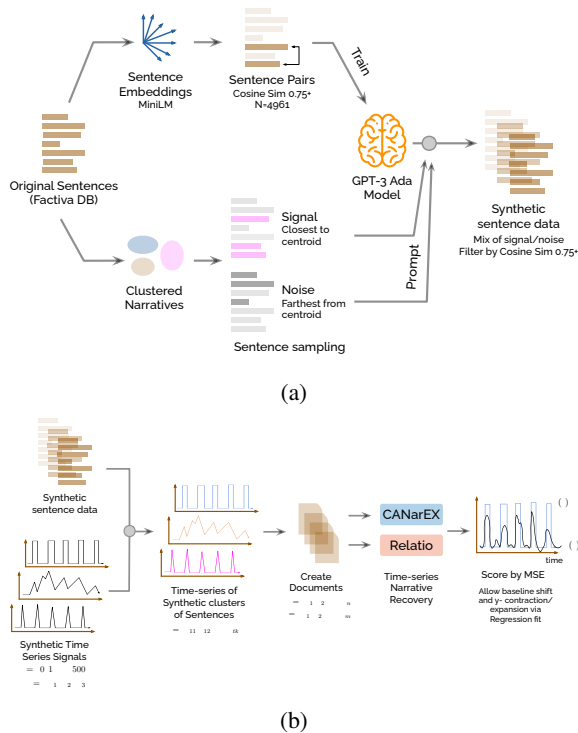
(a)

(b)

Figure 5: **Evaluation pipeline** (a) Fine-tune GPT-3 using similar sentences and sample synthetic data from GPT-3 (b) Generating synthetic time-series cluster sentences and evaluating the pipeline.

and noise sentences where chosen farthest from the cluster centroids. In case of noise, there could be a scenario of overlapping clusters, but since the points were retrieved from clusters not associated with the chosen 3 topic clusters, the impact is assumed to be minimal.

We use these test sentences to generate synthetic sentences or similar sentences using the previously fine-tuned GPT-3, Figure 6 (b). We rerun paraphrase mining (Reimers and Gurevych, 2020) and set similar threshold ($0.75 \leq$ cosine score < 0.99) as our train dataset to further clean the generated test dataset. This filters out the noisy data, and acts as data quality verification on output of GPT-3.

We also generate 3 distinct time-series signals: periodic pulse, periodic spike and a typical time-series pattern with random rise and fall. These are then randomly paired with the synthetic sentences (topic clusters sentences) of 3 topics generated using GPT-3. This generates a time-series of synthetic clusters over time. At each time point, the height of signal represents the N number of sentences associated with a cluster C, Figure 7. The sentences are then collated into random documents. The distribution of synthetic test sentences into

documents is similar to the input Factiva's sentence distribution per document. Noise varied randomly between 0.1 to 0.3% of the number of sentences in the document.

We run our CANarEx pipeline and also the Relatio model on these sentences (collated into a random document) to extract narratives. The synthetic test data creation and evaluation pipeline is presented in Figure 5 (b).

Finally, we design synthetic data for 2 scenarios (see Table 2). One with co-reference resolved sentence pairs, and one without co-reference resolution. The latter synthetic data is constructed specifically to test Relatio as it does not do co-reference resolution.

(a)

(b)
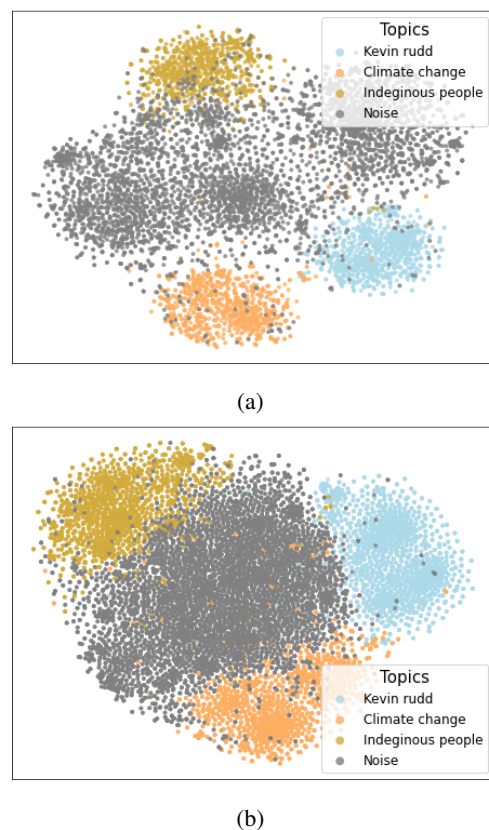
Figure 6: **Test clusters (original data)** (a) Original factiva data (b) GPT3 fine-tuned data with noise
The FIT-SNE (Linderman et al., 2019) approach is used to plot the lower dimensional representation of sentence embeddings of the Factiva dataset.

Evaluation pipeline

The overall idea of evaluation using the regression approach is to validate the number of time-series sentences retained by the model and if it recovers the temporal signature of the original time-series sentences. We are interested in only how

Figure 7: Synthetic time-series clusters (noise combined with cluster labels).

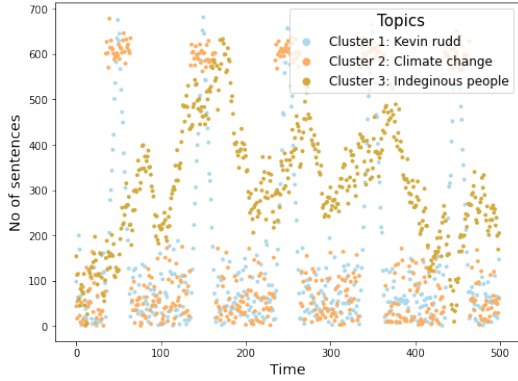|		# sentences	# sampled sentences	
Labels	**Topics**		**# true sentences**	**# co-ref sentences**
1	Kevin Rudd	992	2987	2681
2	Climate change	982	2606	2449
3	Indeginous people	936	2841	2654
4	Noise	4938	8434	8192

Table 2: GPT-3 generated test data

well a respective model recovers the temporal pattern of the synthetic time-series, treating baseline shifts, amplitude expansions and contractions as non-material via regression. The dependent variable is the output of CANarEx or Relatio, i.e. the number of sentences associated with the generated time-series narratives and the independent variable is the number of sentences associated with the input time-series synthetic data. Lower MSE indicates a better recovery method and more strictly defines the accuracy of the technique, where R_sq might spuriously reward general correlation whilst under-representing inaccuracies due to sharp movements of the synthetic time-series.

We also do A/B testing and compare precision/recall scores for the recovery of the synthetic time-series sentences belonging to each cluster C1-C3.

Baseline model

We ran Relatio (baseline) with default settings on our synthetic narrative time-series clusters. The dimensionality reduction clustering for entities was evaluated for cluster counts 5, 10, 20, 30, 40, 50 and 100. We re-run clustering as downstream task on the dimensionality reduced narratives to recover the time-series narratives. We report the MSE results of this in comparison with our model. We report the comparison for 2 of the entity cluster

counts of Relatio: 40 (best entity mapping as determined through manual audit) and 100 (default Relatio entity cluster count). We also compare our model to a null model. We generate the null model by randomly assigning cluster numbers to sentences and evaluating the MSE of such a model.

The results are presented in Table 3. The performance is tabulated per model variant, and per synthetic cluster for each model variant. The performance of CANarEx is better than both Relatio and the null model for each of the three clusters (lowest MSE across clusters). CANarEx uncovered 4 clusters, and Relatio identified 9 (for entity cluster 40) and 6 (for entity cluster 100) clusters. The MSE are presented for recovered clusters that have the lowest score in comparison to the original synthetic clusters. The recovery is presented in Figure 8 for the CANarEx model. A similar analysis with the inclusion of the filtering step is documented in Appendix B. Most of the noise was categorised into a separate cluster.

A/B test on percentage of sentences recovered across clusters between CANarEx and Relatio (t-test one-sided): +6.943 (p-value: 0.001). 95% of the time, CANarEx recovers at least 26.8% more sentences than Relatio. The precision/recall scores are provided in Table 4 for the recovery of the synthetic time-series sentences belonging to each cluster C1-C3.
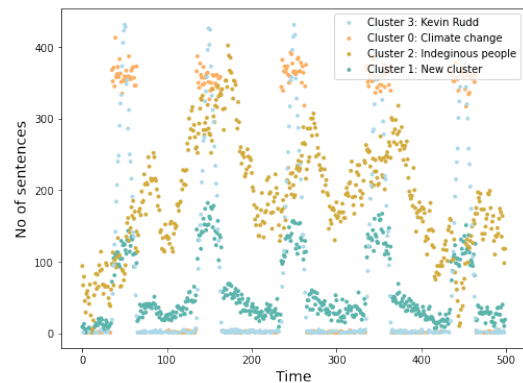

Figure 8: Synthetic time-series clusters recovered by CANarEx (without noise).

Ablation studies

Given the multi-component nature of our framework, we also evaluated the performance through an ablation study. The ablation study compared four configurations of CANarEx: with and without co-reference resolution, and with and without micro-narratives generation.

	CANarEx	Relatio (40 entity clusters)	Relatio (100 entity clusters)	Null model (with co-referencing)	Null model (without co-referencing)
Original Cluster			MSE		
1	46.6	378.2	471.8	5422.7	4792.7
2	79.4	6818	1444.9	8600.6	7664.6
3	155.4	555.5	580.7	9943.7	8785.6

Table 3: Evaluation results of CANarEx with baseline model (Relatio) and null model (full table in Appendix A, A1).

	CANarEx			Relatio (40 entity clusters)			Relatio (100 entity clusters)		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Cluster 1	0.94	0.89	0.91	0.85	0.62	0.71	0.7	0.53	0.6
Cluster 2	0.97	0.83	0.89	0.31	0.62	0.41	0.28	0.39	0.33
Cluster 3	0.97	0.85	0.9	0.96	0.84	0.9	0.94	0.74	0.83
Cluster 4	0.71	0.94	0.81	0.41	0.42	0.41	0.28	0.31	0.3

Table 4: Precision/Recall scores for the recovery of the synthetic time-series sentences belonging to each cluster. Cluster 4 here is assigned to noise.

Table 5 presents the evaluation of CANarEx for all extracted narratives. This evaluation does not include the optional filtering step, and exercises the framework for the four combinations of co-reference resolution and micro-narratives generation. The performance of the framework when both co-reference resolution and micro-narratives generation are used is marginally worse than the best performing configuration (co-reference resolution only). The performance is markedly worse when co-reference resolution is not used. This is understandable as removing co-reference resolution removes all the subsequent mentions of an entity and therefore the extracted narratives are not contextual. Turning off co-reference resolution also has a cascading effect on the filtering mechanisms. Ablation studies of the optional filtering step are discussed in Appendix D.

The generation of micro-narratives increases the error rate, but only marginally. We expect this to be a consequence of the narrative splitting step introducing a degree of noise in the resulting set of micro-narratives. Secondly, the test data synthesis is geared to associate each sentence with only one cluster. In contrast, the CANarEx framework identifies more than one micro-narrative in some of the test sentences, and these micro-narratives can belong to different clusters, therefore penalizing MSE. Despite these observations, the micro-narratives still retain the concepts, as demonstrated via the embedding space (see Figure 9). Therefore, the marginal increase in the error rate is an

|		Original Cluster (MSE)		
Co-ref	**Micro-EVEs**	**1**	**2**	**3**
✓	✓	46.6	79.4	155.4
✓	✗	45	74.7	151.4
✗	✓	181.3	135.3	312.8
✗	✗	152.9	126	302.4

Table 5: CANarEx performance over all of the extracted narratives from the synthetic narrative time-series data (full table in Appendix A, A2).

acceptable trade-off given the improved readability and recovery enabled through micro-narratives generation.

Qualitative analysis

We qualitatively evaluate the performance of CANarEx by analysing the micro-narratives it extracted for the topic of indigenous people from the synthetic text corpus. In particular, we examine the micro-narratives for the entity 'indigenous voice' within this corpus, as it is one of the central entities of the topic and is associated with multiple themes and narratives. Results of the extraction with TopN document level filtering are presented Figure 10 (Appendix E for results of CANarEx without filtering and with Textrank filtering). As evident, the extracted micro-narratives are all aligned with the public discourse on indigenous people in Australia. The verbs identified with indigenous voice are *enshrine, have, need, provide, will be, would be* and are linked with strong narratives like *have strong national voice*, *have legislative voice*, and *enshrine (in the) constitution*. The analysis also re-
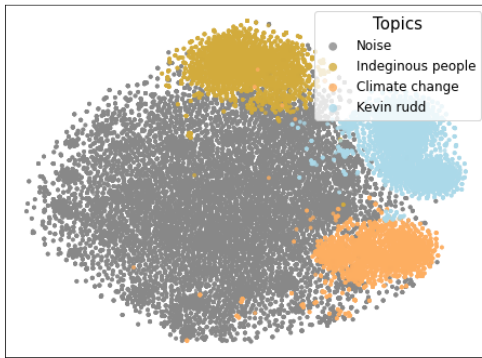
(a)
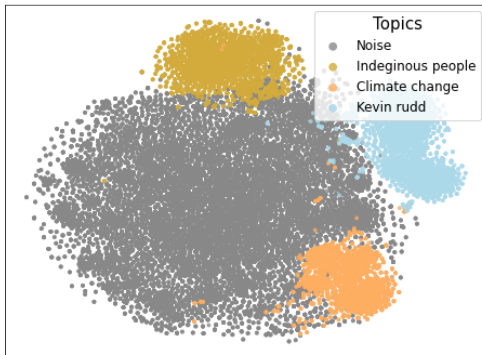
(b)

Figure 9: **Test clusters (synthetic data)** (a) Without micro-narratives (b) With micro-narratives

veals the need for better span management, as motivated by the result "e1:indigenous voice - v:will be - e2:body" wherein the word *body* should have been *standalone body*.

5 Conclusion

In this work we introduced the CANarEx framework for contextually aware narrative extraction for text-as-data applications. The contextual aspects of the narrative are retained via the co-reference resolution step and by the use of the complete E-V-Es triplets for clustering. We also present an optional step that enables filtering of the extracted micro-narratives. We demonstrate the framework on two datasets - parliamentary proceedings and news articles. Further, we generate synthetic test data using a novel GPT-3 based technique for evaluation of the framework, and demonstrate better performance than the baseline model. The results are reinforced through an ablation study exercising the various components of the pipeline, and a qualitative study demonstrating the applicability of the framework.
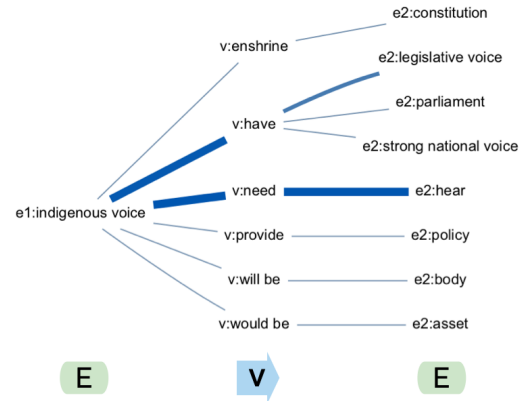

Figure 10: CANarEx generated micro-narratives on topic of indigenous voice on synthetic test data (filtered via TopN document level clustering).

Limitations

The CANarEx framework's performance is limited by the underlying models that it leverages for co-reference resolution and SRL extraction, although we note that these models can be substituted with improved variants when available. Further evaluation is required to understand the performance of SRL extraction on corpuses enhanced with co-reference resolution. An option would be to perform co-reference resolution and SRL simultaneously. SRL extraction can generate E-V-Es that are not narratives, and the generation of micro-narratives introduces a degree of noise as well. We mitigate this through filtering and stop word removal, but a qualitative analysis of the resulting micro-narratives is an avenue for improvement. Finally, the clustering approach that we employ to reduce the number of E-V-Es to a viable subset can also remove valid E-V-Es through overtly aggressive filtering. While there is some penalty from clustering only the top sentences, Textrank performs better than document level clustering and the improvement is correlated with the ratio of sentences extracted.

Acknowledgements

References

Elliott Ash, Germain Gauthier, and Philine Widmer. 2021. Text semantics capture political and economic narratives. *arXiv preprint arXiv:2108.01720*.

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the similarity function of textrank for automated summarization. *CoRR*, abs/1602.03606.

Janet M. Box-Steffensmeier, John R. Freeman, Matthew P. Hitt, and Jon C. W. Pevehouse. 2014. *Time Series Analysis for the Social Sciences*. Analytical Methods for Social Research. Cambridge University Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-Semantic Parsing. *Computational Linguistics*, 40(1):9–56.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. *Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

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

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. 2019. Text as Data. *Journal of Economic Literature*, 57(3):535–74.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada. Association for Computational Linguistics.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019a. Span-BERT: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*.

Mandar Joshi, Omer Levy, Daniel S. Weld, and Luke Zettlemoyer. 2019b. BERT for coreference resolution: Baselines and analysis. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Ashkan Kazemi, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Biased textrank: Unsupervised graph-based content extraction. *ArXiv*, abs/2011.01026.

George C. Linderman, Manas Rachh, Jeremy G. Hoskins, Stefan Steinerberger, and Yuval Kluger. 2019. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature Methods*, 16(3):243–245.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marco Rospocher, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. 2016. Building event-centric knowledge graphs from news. *Journal of Web Semantics*, 37-38:132–151.

Peng Shi and Jimmy Lin. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *ArXiv*, abs/1904.05255.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*.

Evangelia Spiliopoulou, Eduard Hovy, and Teruko Mitamura. 2017. Event detection using frame-semantic parser. In *Proceedings of the Events and Stories in the News Workshop*, pages 15–20, Vancouver, Canada. Association for Computational Linguistics.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Sandra Wankmüller. 2021. Neural transfer learning with transformers for social science text analysis. *CoRR*, abs/2102.02111.

Hao Zhang, Frank Boons, and Riza Batista-Navarro. 2019. Whose story is it anyway? Automatic extraction of accounts from news articles. *Information Processing & Management*, 56(5):1837–1848.

A   Appendix: Evaluation results of CANarEx and Ablation

Tables A1 and A2 show the full result set of Tables 3 and 5 respectively.

B   Appendix: Evaluation results with the filtering step

Tables B1 and B2 show the performance of CANarEx after applying the two filtering techniques (document level clustering and Textrank filtering techniques respectively). We note that the reduction of E-V-Es resulting from these techniques does penalize the MSE of CANarEx, although the performance continues to be better than the baseline model.

CANarEx					
Original Cluster	Best Match Cluster	Intercept	Coef	MSE	R_sq
1	3	-3.8	1.3	46.8	0.998
2	0	-3.8	1.4	79.4	0.998
3	2	-4.5	1.3	155.4	0.986
Relatio (40 entity clusters)					
1	3	-7.3	3.5	378.2	0.982
2	7	-134.7	6.4	6818	0.828
3	0	7.8	2.3	555.5	0.94
Relatio (100 entity clusters)					
1	5	-19.2	3.2	471.8	0.977
2	4	-43.7	5.8	1444.9	0.968
3	2	4.5	2.2	580.7	0.937
Null model (with co-referencing)					
1	2	-139.5	1.3	5422.7	0.73
2	1	-187.2	2	8600.6	0.8
3	4	202.7	0.3	9943.7	-8.009
Null model (without co-referencing)					
1	3	-132.6	1.3	4792.7	0.708
2	3	-184.9	2	7664.6	0.802
3	2	188.9	0.3	8785.6	-7.699

Table A1: Evaluation results of CANarEx with baseline model (Relatio) and null model

C   Appendix: Factiva result example

The result of Textrank filtering on the topic of crime within the 'First nations' Factiva data is shown in Figures C1. The links between the E-V-Es are weighted by frequency of occurrence. This enables us to identify the dominant micro-narratives per entity pair, such as the higher occurrence of *e1:mandatory sentence - v:produce - e2:result* compared to other micro-narratives.
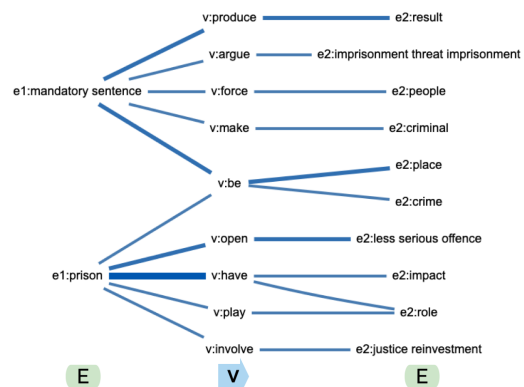

Figure C1: CANarEx generated micro-narratives on topic of criminal justice, derived from the Factiva dataset.

All Clusters							
Co-ref	Micro-EVEs	Original Cluster	Best match Cluster	Intercept	Coef	MSE	R_sq
✓	✓	1	3	-3.8	1.3	46.8	0.998
		2	0	-3.8	1.4	79.4	0.998
		3	2	-4.5	1.3	155.4	0.986
✓	✗	1	0	-4.2	1.2	45	0.998
		2	3	-5.1	1.3	74.7	0.999
		3	2	-4.2	1.3	151.4	0.986
✗	✓	1	1	-3.8	2.2	181.3	0.991
		2	2	-4	1.7	135.3	0.997
		3	0	-2	1.6	312.8	0.967
✗	✗	1	0	-6.1	2	152.9	0.993
		2	2	-4.4	1.6	126	0.997
		3	3	-4	1.5	302.4	0.968

Table A2: CANarEx performance over all of the extracted narratives from the synthetic narrative time-series data.

CANarEx (TextRank)				
Original Cluster	Best Match Cluster	Intercept	Coef	MSE	R_sq
1	0	-2.5	6.1	345.1	0.986
2	1	-0.8	5.7	444.1	0.991
3	3	12.5	5.3	677.4	0.935
Relatio (40 clusters)					
1	4	-1.6	5.4	468.5	0.977
2	2	-26.1	8.7	1952.6	0.956
3	3	20.1	3.7	759.6	0.916
Relatio (100 clusters)					
1	5	2.1	12.3	1564	0.92
2	0	-14.7	15.3	2829.4	0.935
3	1	34.4	6.9	1428.2	0.829

Table B1: Evaluation results of CANarEx with Textrank filtering, compared to baseline model (Relatio)

D   Appendix: Ablation study of the filtering step

Continuing the ablation study with the inclusion of the optional filtering step, we observe that the same pattern observed in the main study applies here as well (see Table D1 for filtering through Textrank clustering, and Table D2 for filtering through TopN document level clustering). Finally, comparing along the dimension of filtering mechanisms, Textrank performs better than TopN document level clustering as the latter clips narratives aggressively.

E   Appendix: Qualitative study

The unfiltered set of micro-narratives is large and the link between the E-V-Es is more diffuse compared to the filtered sets. The filtering process has resulted in a clearer association between the entities of the unpruned E-V-Es.

Figure E1: CANarEx generated micro-narratives on topic of indigenous voice on synthetic test data (unfiltered result set).

CANarEx (TopN)

Original Cluster	Best Match Cluster	Intercept	Coef	MSE	R_sq
1	1	-5.3	28.5	782.4	0.968
2	0	2.7	30.3	1530.6	0.969
3	2	-15.5	28.3	2359.2	0.728
Relatio (40 clusters)					
1	3	2.5	45.9	2837.2	0.845
2	2	-14.8	59.9	8218.8	0.785
3	1	12.6	28.2	2393.2	0.677
Relatio (100 clusters)					
1	3	-0.1	43.7	2482.5	0.867
2	2	-17.8	53.6	9196.4	0.753
3	1	15	27.3	2454.1	0.666

Table B2: Evaluation results of CANarEx with TopN document level filtering, compared to baseline model (Relatio)
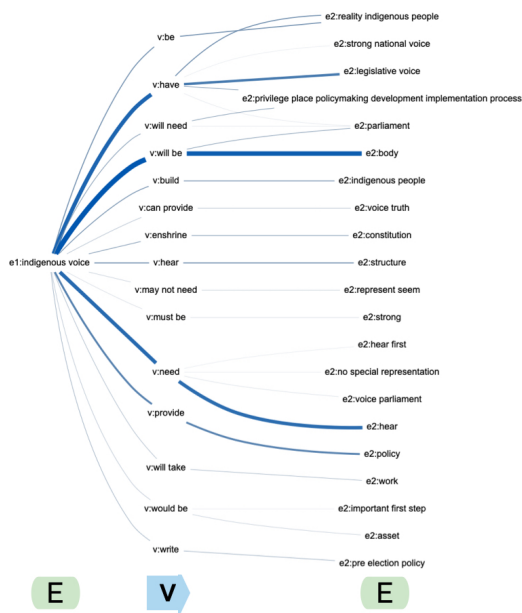

Figure E2: CANarEx generated micro-narratives on topic of indigenous voice on synthetic test data (filtered via Textrank).

			Textrank				
Co-ref	Micro-EVEs	Original Cluster	Best match Cluster	Intercept	Coef	MSE	R_sq
✓	✓	1	0	-2.5	6.1	345.1	0.986
		2	1	-0.8	5.7	444.1	0.991
		3	3	12.5	5.3	677.4	0.935
✓	✗	1	2	-2.2	6	313.2	0.988
		2	3	-0.1	5.5	389.7	0.992
		3	1	12.3	5.2	618.6	0.941
✗	✓	1	1	7	32.6	2454.9	0.869
		2	0	5	26.5	2093.3	0.953
		3	3	81.1	20.9	3511.2	0.441
✗	✗	1	2	5.2	30.2	1977	0.897
		2	1	3.9	25.7	1916.8	0.957
		3	3	73.4	20.9	3188	0.518

Table D1: CANarEx performance over the extracted narratives filtered through Textrank clustering.

			TopN				
Co-ref	Micro-EVEs	Original Cluster	Best match Cluster	Intercept	Coef	MSE	R_sq
✓	✓	1	1	-5.3	28.5	782.4	0.968
		2	0	2.7	30.3	1530.6	0.969
		3	2	-15.5	28.3	2359.2	0.728
✓	✗	1	0	-5.2	28.3	775.6	0.969
		2	2	2.6	30.1	1529.2	0.97
		3	1	-13.4	27.9	2422.4	0.719
✗	✓	1	3	-0.5	37.1	1641.2	0.916
		2	2	2.1	30.7	1580.8	0.965
		3	1	3.6	27.6	2413.9	0.673
✗	✗	1	0	-1.7	35.8	1686.2	0.913
		2	2	0.3	30.3	1585.2	0.965
		3	1	-2	27.4	2324.3	0.689

Table D2: CANarEx performance over the extracted narratives filtered through TopN document level clustering.