Constructing and Interpreting Causal Knowledge Graphs from News

AAAI 2023 Summer Symposium, AI4FinTech, 18 – 19 July 2023

Fiona Anting Tan, Debdeep Paul,

Sahim Yamaura, Miura Koji, See-Kiong Ng

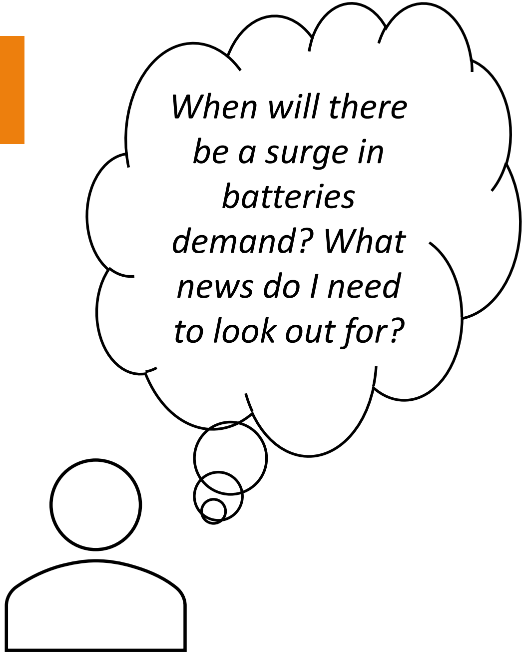
https://arxiv.org/abs/2305.09359

© Copyright National University of Singapore. All Rights Reserved.

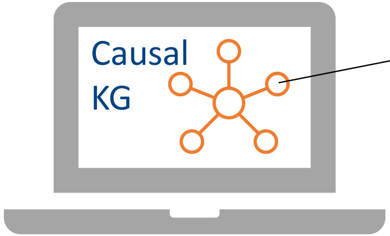
Understanding causal relationships in news helps users make informed decisions.

- Many jobs rely on news to learn about causal events in the past and present, to make informed decisions and predictions about the future.
- Such careers require stakeholders to have a general understanding of how one thing leads to another.
- How can we manage the **large volume of news** in an **automated** manner that **helps users understand and interpret causal events**?

© Copyright National University of Singapore. All Rights Reserved.



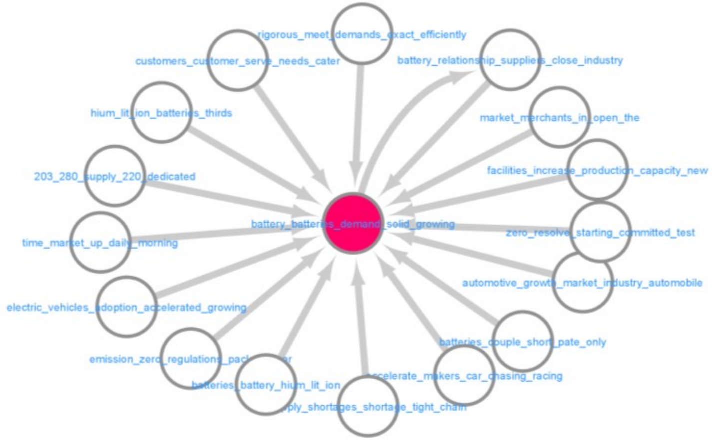
Parts Purchaser for EV Batteries



1. Purchaser can search for relevant nodes to understand about the factors that affect battery demand from KG

2. Purchaser can set these *Cause* topics as new alerts

Search Node: battery_batteries_demand_solid_growing

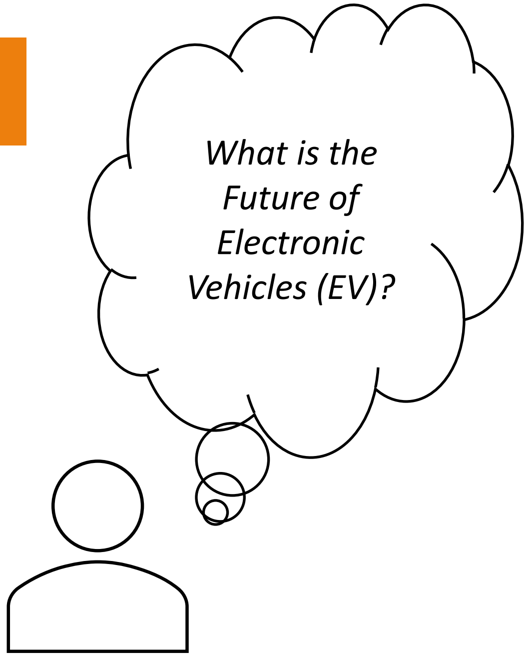


Identified Causes:

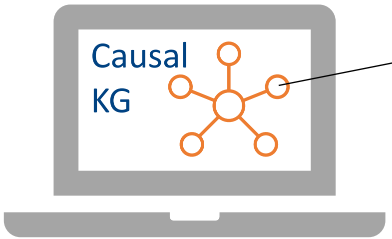
- *fuel_environmental_concerns_increasing_costs:* The increasing high fuel costs and environmental concerns
- *battery_relationship_suppliers_close_industry:* together with the TESLA project … promoting as part of the European battery projects
- *electric_vehicles_adoption_accelerated_growing:* the electric vehicle revolution

…

© Copyright National University of Singapore. All Rights Reserved.



Strategy Consultant



Actively explores causal relations leading to and from topics about EV

Drivers of EV Growth:

- Potential to save money and lower pollution
- Government decisions
  - Several governments now provide subsidies for the use of environment - friendly vehicles
  - CO_2 Emissions target in Europe
  - China aiming to keep a key market growing as the broader economy slows
  - Singapore has announced ambitious targets for 2040 when we plan to phase out internal combustion engine vehicles
- New investments and advancements in technology; Billions spent on R&D

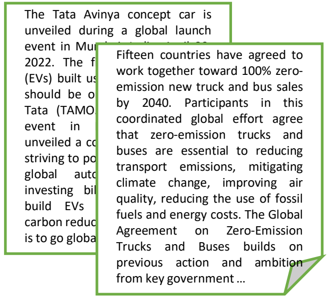
Consequences of EV Growth:

- Increase in battery charging stations and battery backers
  - Greater efforts are needed to roll out enough charging infrastructure
  - Already struggling to keep up with strong demand for lithium - ion batteries
- More partnerships & joint ventures in this arena
  - Automobile giants like Tesla, Tata, Maruti and Hyundai have already established several partnerships
- Growth of the luxury SUV market in north America
- Tesla over took Toyota as the word's most valuable auto mater in July; Tesla is more valuable than the next six auto makers combined

…

© Copyright National University of Singapore. All Rights Reserved.

We extract, cluster and represent of causal relations in order to construct our knowledge graph.



News Articles

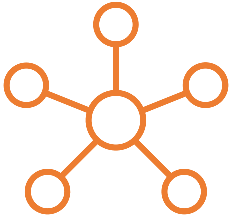
Cause

the chip crisis and the surge in Covid cases in some parts of the world

Effect

Global car manufacturers like Toyota Motor Corp and Ford Motor Corp have decided to slash production sharply

Causal Relations



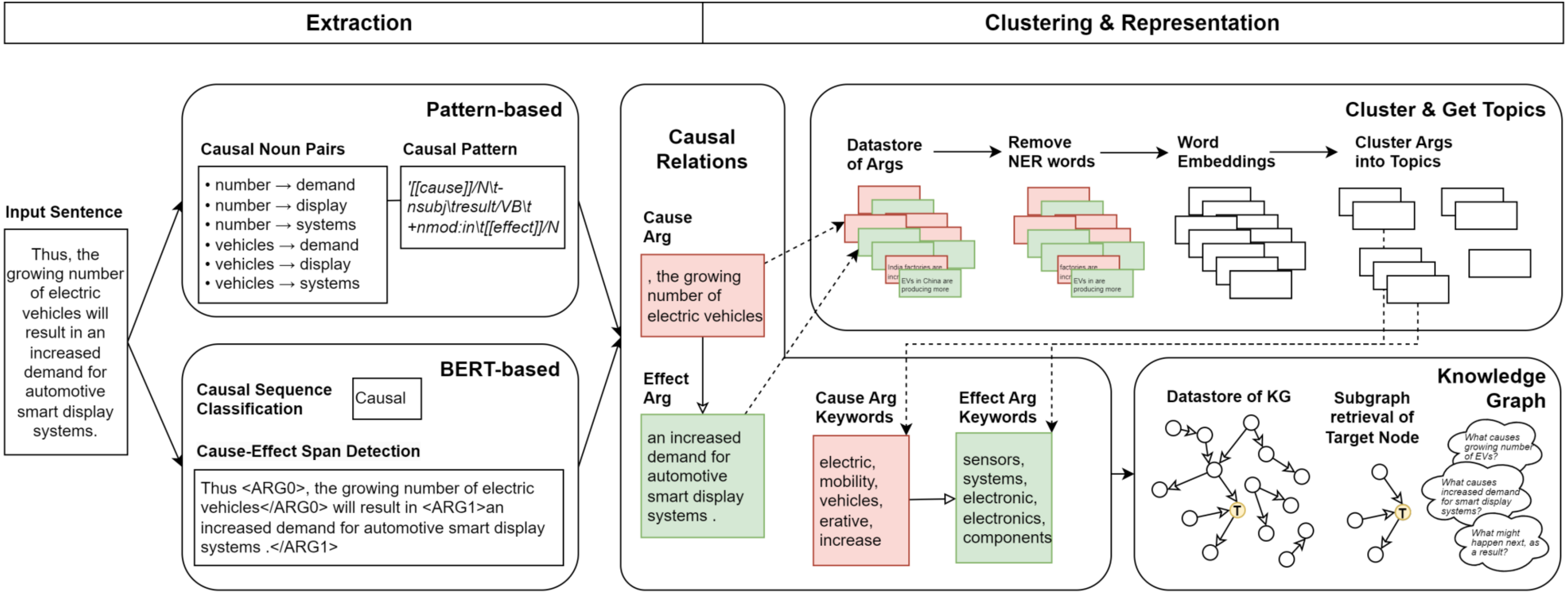
Causal Knowledge Graph

Extraction

Clustering & Representation

© Copyright National University of Singapore. All Rights Reserved.

We extract, cluster and represent of causal relations in order to construct our knowledge graph.



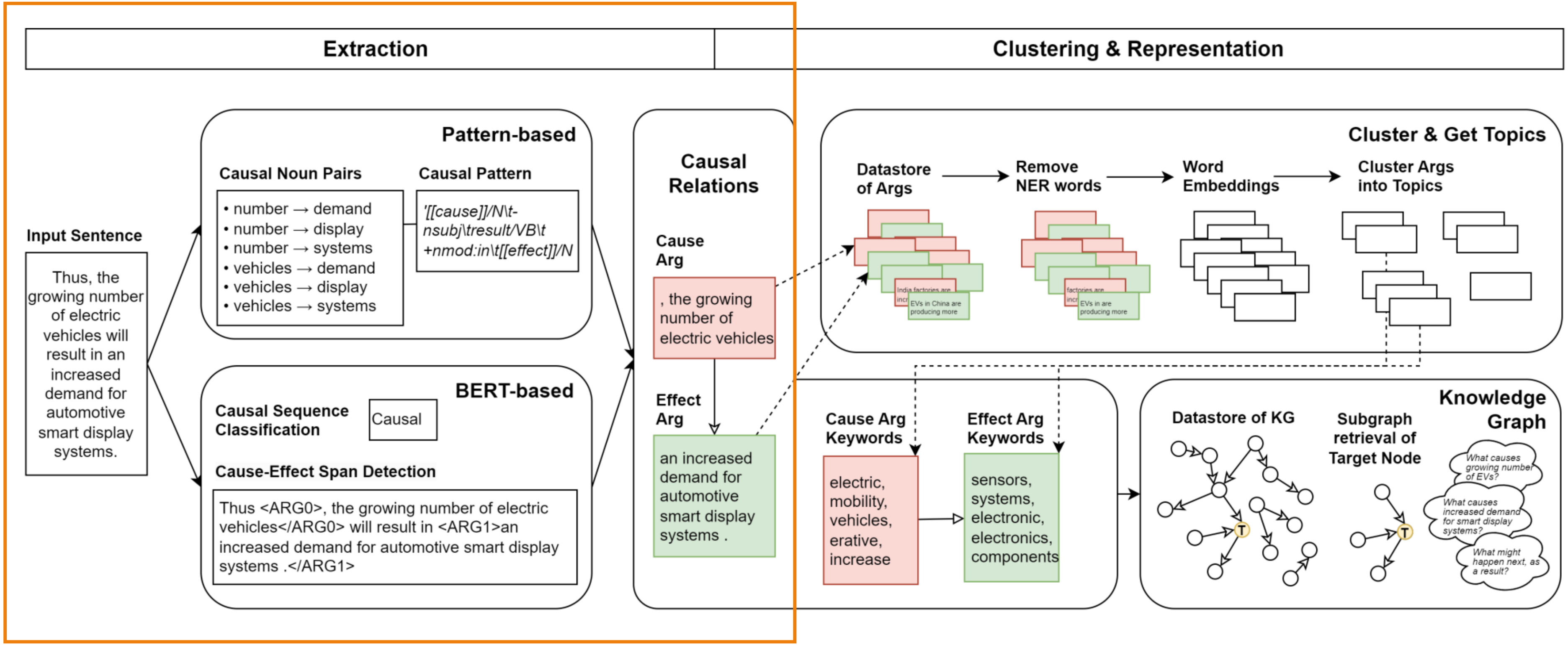
© Copyright National University of Singapore. All Rights Reserved.

We focused on news from the electronics and supply-chain industry.

- **Focus:** Electronics and supply-chain industry news
- **Size:** 6,384 article summaries and 62,151 sentences
- **Time period:** 2017 and 2022
- **Source:** Google News

© Copyright National University of Singapore. All Rights Reserved.

Causal relations were extracted using two methods: (1) Pattern-based and (2) BERT-based.



© Copyright National University of Singapore. All Rights Reserved.

Pattern-based extraction uses linguistic patterns to identify causal relations, based on CauseNet (Heindorf et al., 2020).

Rank	Linguistic Pattern	|Rels|
(A) Directly ($n_{patterns} = 53$)		
1	[[cause]]/N -nsubj cause/VB +dobj [[effect]]/N	904,385
2	[[cause]]/N -nmod:with associated/VBN -acl [[effect]]/N	892,908
3	[[cause]]/N -nsubj lead/VB +nmod:to [[effect]]/N	783,860
(B) Reverse-Engineered ($n_{patterns} = 477$)		
1	[[cause]]/N -nsubj led/VBD +nmod:to [[effect]]/N	67,255
2	[[cause]]/N -nsubj produced/VBD +obj [[effect]]/N	43,188
3	[[cause]]/N -nsubj brought/VBD +obj [[effect]]/N	35,065

Table 1. Common linguistic patterns used.

- Identify shortest dependency path between two nouns
- Identify as causal based on a list of causal linguistic patterns

Original	Pattern-based Extraction	Post-processing	Pre-processing for Clustering
... implementing a furlough scheme aimed at mitigating the impact of a fall in output brought on by a global chip shortage.	• shortage → impact • shortage → fall • shortage → output Pattern: '[[cause]]/N -nmod:by brought/VBN +nmod:of [[effect]]'	... implementing a furlough scheme aimed at mitigating the <ARG1>impact of a fall in output</ARG1> brought on by a global chip <ARG0>shortage </ARG0>.'	... implementing a furlough scheme aimed at mitigating the <ARG1>impact of a fall in output</ARG1> brought <ARG0>on by a global chip shortage</ARG0>.'

Table 2. Processing of pattern-based extracted causal relations.

© Copyright National University of Singapore. All Rights Reserved.

BERT-based extraction uses pre-trained models to get causal information, based on UniCausal (Tan et al., 2023).

- Pre-trained models trained on **Causal Sentence Classification (CSC)**, **Cause-Effect Span Detection (CESD)** and **Causal Pair Classification (CPC)** tasks on external datasets
  - *Causal* sentences based on CSC: Takes arguments from CESD + post-processing
  - *Non-causal* sentences based on CSC: If sentence contains only one pair of arguments and if CPC finds them to be *Causal*, then these arguments are accepted.

Table 3. Processing of BERT-based causal relations.

Original	BERT-Based Extraction		Post-Processing	
	CSC	CESD	Final	Method
Ford is shutting its car factories in India after Ford India racked up more than $2bn in losses over the past decade.	0	<ARG1>Ford is shutting its car factories in India</ARG1> after <ARG0>Ford India racked up more than $2bn in losses over the past decade.</ARG0>	<ARG1>Ford is shutting its car factories in India</ARG1> after <ARG0>Ford India racked up more than $2bn in losses over the past decade.</ARG0>	Add Causal Relations based on CPC

© Copyright National U

It is important to focus on both precision-driven pattern-based and recall-driven BERT-based approaches.

- Pattern-based approach: High precision, low recall
- BERT-based approach: High recall, low precision

Method	Extraction			
	$\lvert Sents \rvert$	$\lvert Sents \rvert$	$\lvert Rels \rvert$	Avg Rel Support
Pattern-based	1,006	611	975	1.032
BERT-based	19,250	15,702	19,192	1.003
Total	20,255	15,902	20,086	1.008

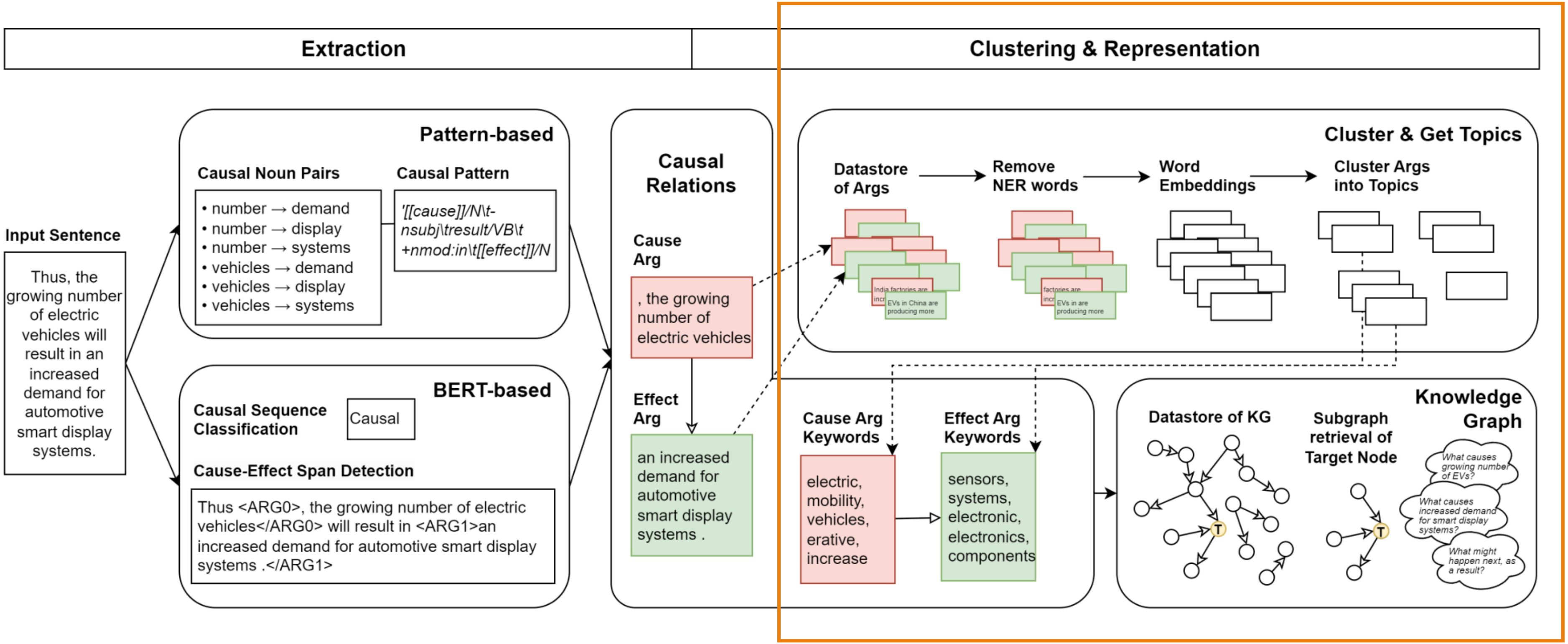
Table 4. Summary Statistics from Extraction step.

Extraction Method	P	R	F1
Pattern-based	**100.00**	4.08	7.84
BERT-based	76.09	71.43	73.68
Both	75.00	**73.47**	**74.23**

Table 5. Performance metrics for extraction. Scores reported in percentages (%). Top score per column is in bold.

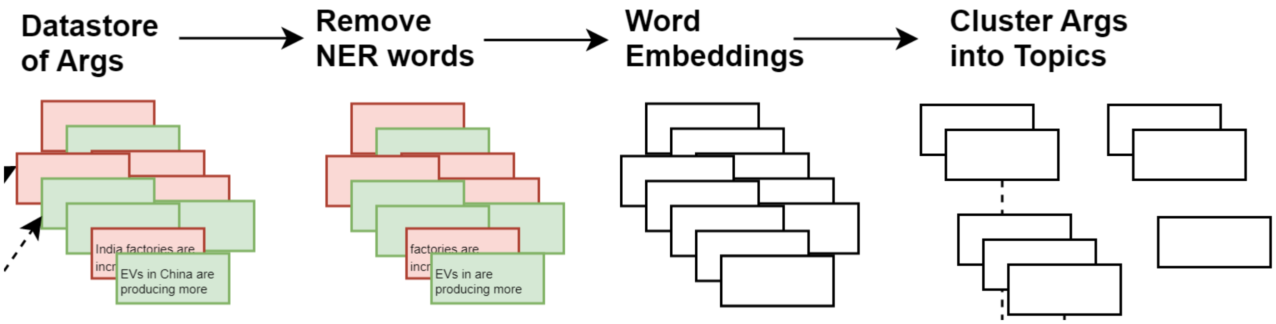
© Copyright National University of Singapore. All Rights Reserved.

Causal relations were extracted using two methods: (1) Pattern-based and (2) BERT-based.



© Copyright National University of Singapore. All Rights Reserved.

We wish to cluster the arguments that have similar meaning together into the same topic/ node.



1. Remove named-entities from span (Stanford NER Tagger – Finkel et al., 2005)
2. Generate BERT-based word embeddings from each span (SimCSE – Gao et al., 2021)
3. Cluster spans using K-Means

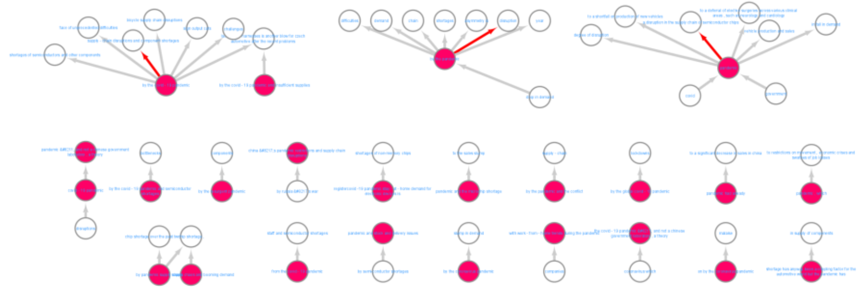
© Copyright National University of Singapore. All Rights Reserved.

Clustering helps to create denser graphs that allows for more meaningful causal relationships to be learnt.

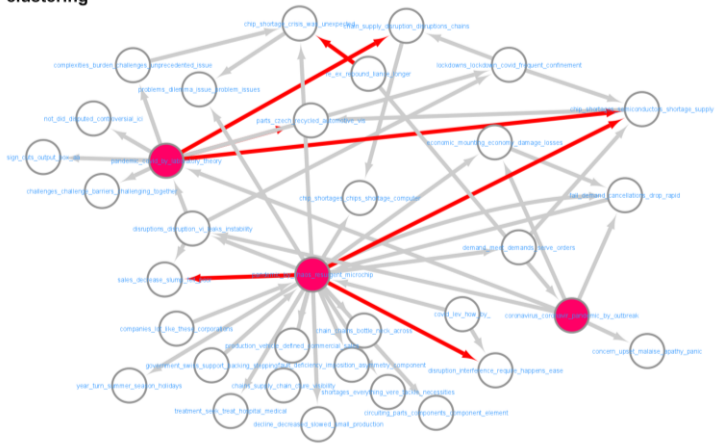
	Before Clustering	After Clustering
No. of Unique Nodes, $\lvert V \rvert$	35,230	3,000
No. of Unique Edges, $\lvert E \rvert$	20,086	17,801
Total Weight, $\sum s$	20,254	19,965
No. of Subgraphs	15,686	1
Avg Clustering Coefficient	$9.81e^{-06}$	$1.75e^{-02}$
Avg Degree Centrality	$3.24e^{-05}$	$3.96e^{-03}$
Avg Eigenvector Centrality	$6.64e^{-05}$	$1.32e^{-02}$
Transitivity	$4.17e^{-04}$	$8.81e^{-03}$

Table 6. Graph summary statistics.

I. Before clustering



II. After clustering



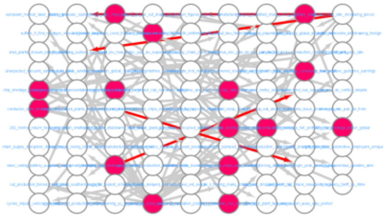
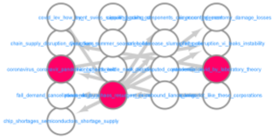
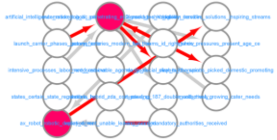
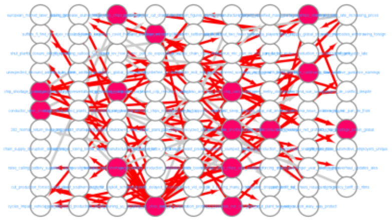
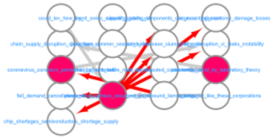
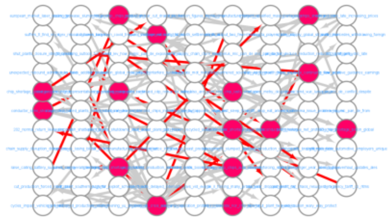
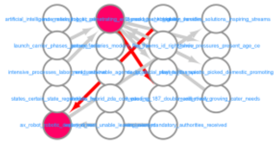
© Copyright National University of Singapore. All Rights Reserved.

Applications

- Summarization:
  - Collect relations in news across time
  - Search and monitor on-going trends in news
- Answering causal questions and predicting future events
- Inferring transitive causal relations
- Trend monitoring

© Copyright National University of Singapore. All Rights Reserved.

Our KG can be helpful for monitoring heated causal topics and news trends across time.

Article date \ Target node	"chip shortage"	"pandemic"	"robotics"
Before 2020			
2020 - 2021			
After 2021			

© Copyright National U

Conclusion

- For extraction, it is important to focus on both precision-driven pattern-based and recall-driven BERT-based approaches.
- Clustering creates denser graphs that allows for meaningful causal relationships to be learnt (E.g. causal chains).
- Our work can be replicated onto many other domains and has various real-life applications in the workplace.
- Future Directions:
  - Readability of nodes/ topics
  - Expand coverage

© Copyright National University of Singapore. All Rights Reserved.

References

Heindorf, S.; Scholten, Y.; Wachsmuth, H.; Ngomo, A. N.; and Potthast, M. 2020. CauseNet: Towards a Causality Graph Extracted from the Web. In d'Aquin, M.; Dietze, S.; Hauff, C.; Curry, E.; and Cudré-Mauroux, P., eds., *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, 3023–3030. ACM.

Tan, F. A.; Zuo, X.; and Ng, S.-K. 2023. UniCausal: Unified Benchmark and Repository for Causal Text Mining. In *Big Data Analytics and Knowledge Discovery*. Springer International Publishing.

Finkel, J. R.; Grenager, T.; and Manning, C. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 363–370. Ann Arbor, Michigan: Association for Computational Linguistics.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6894–6910. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

© Copyright National University of Singapore. All Rights Reserved.

Thank you.

- Link to Preprint: https://arxiv.org/abs/2305.09359
- For questions/ feedback, feel free to contact us:

Fiona Anting Tan
Institute of Data Science
National University of Singapore, Singapore
tan.f@u.nus.edu

© Copyright National University of Singapore. All Rights Reserved.