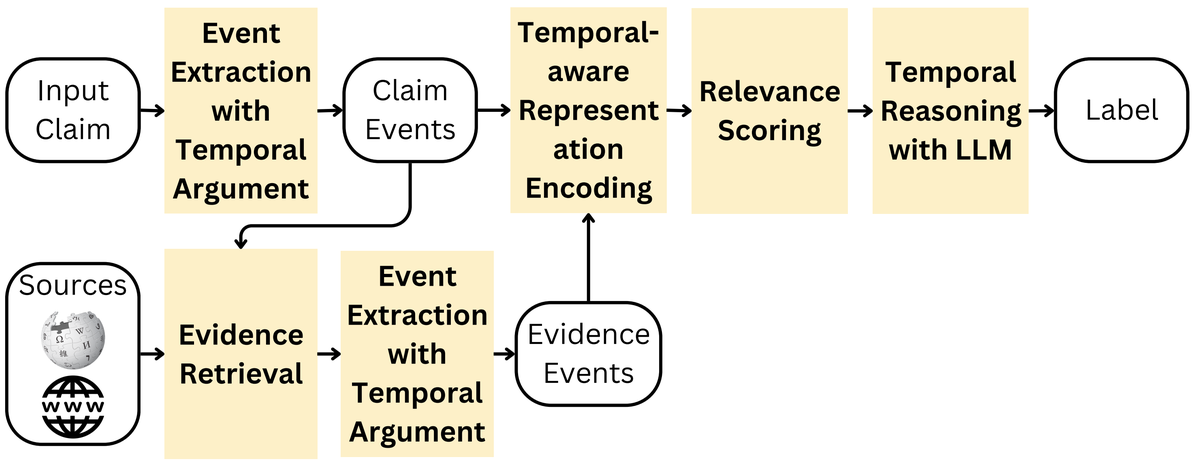
Time Matters: An End-to-End Solution for Temporal Claim Verification

Anab Maulana Barik , Wynne Hsu, Mong Li Lee
School of Computing, Institute of Data Science, Centre for Trusted Internet & Community
National University of Singapore

Motivation

- Automated claim verification plays an essential role in fostering trust in the digital space.
- Temporal claim verification brings new challenges where cues of the temporal information need to be extracted, and temporal reasoning involving various temporal aspects of the text must be applied.
- Therefore, we propose TACV, an end-to-end solution for temporal claim verification that considers the temporal information in claims to obtain relevant evidence sentences and harnesses the power of a large language model for temporal reasoning.

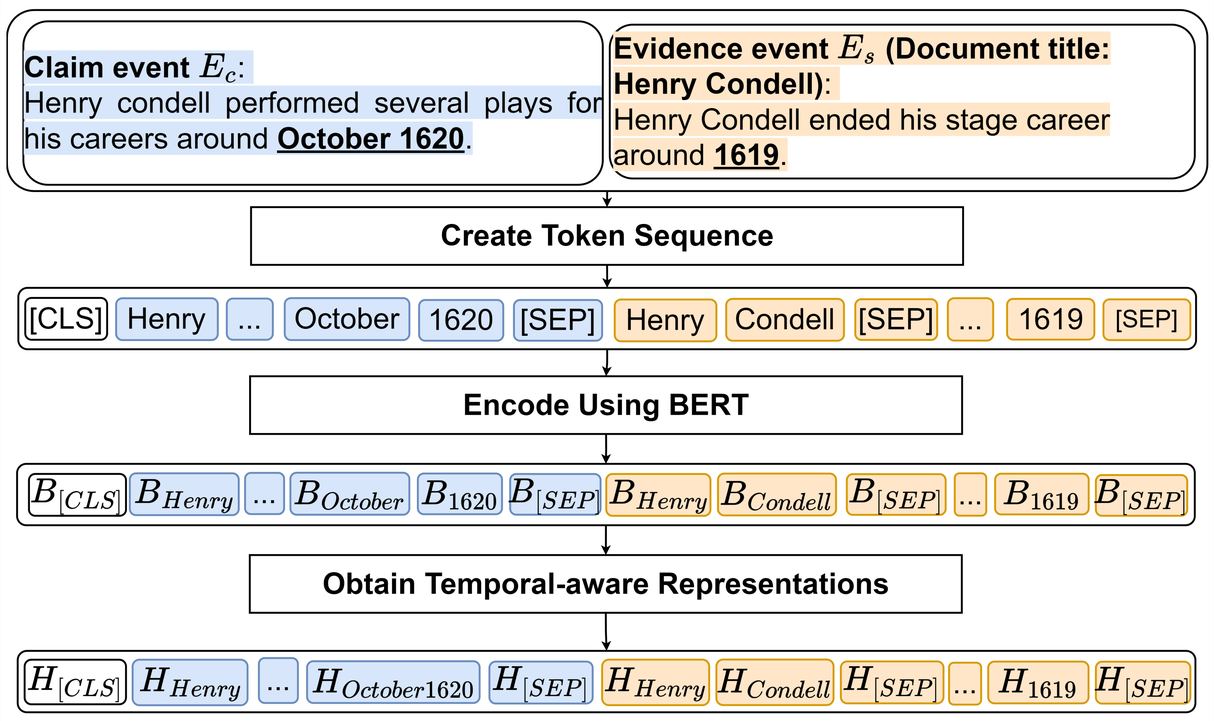
Temporal Aware Claim Verification Framework



Step 1: Event Extraction with Temporal Arguments

- An event comprises of:
  - **Core** information (e.g. who, what, where)
  - **Temporal** information (dates, duration)
- We employ Semantic Role Labelling (SRL) to extract events:
  - **Core** information: concatenation of the predicate and all non-temporal arguments.
  - **Temporal** information: concatenation of the temporal arguments.

Step 2: Temporal-aware Representation Encoding



- We first obtain the contextual representations using BERT.
- We add the representation of the date tokens with temporal encoding and feed them into a transformer to obtain the temporal-aware representation.

Step 3: Relevance Scoring

- We construct an event-level graph where each node is a pair of <claim event, sentence event>
- We apply a Graph Attention Network (GAT) to propagate information among the sentence events.
- We apply a max-wise operation over the final representation of each node and followed by a linear layer to obtain the relevance score of evidence sentence to a claim event.

Step 4: Temporal Reasoning with LLM

- We prompt text-davinci-003 to predict the label of each claim event with the top-k evidence sentences
- The label of the overall claim is determined with:
  - REFUTES: if **ANY** claim event is predicted as REFUTES
  - SUPPORTS: if **ALL** claim events are predicted as SUPPORTS
  - NOT ENOUGH INFO: otherwise

T-FEVER and T-FEVEROUS Datasets

- We curate two temporal claim datasets based on the general claim datasets FEVER, FEVER2.0, and FEVEROUS.
- We identify temporal claims that have at least one temporal argument.
  - **Ordering**: temporal predicate such as “before” or “after”.
  - **Duration**: temporal predicate such as “for 5 years” or “over 3 months”.
- We create new claims by manipulating the temporal arguments of the original claim so that it is either SUPPORTED or REFUTED with the evidence sentences.
- Ground-truth includes the label for the overall claim as well as individual claim events.

	T-FEVER				T-FEVEROUS			
	Single event		Multiple events		Single event		Multiple events	
	Train set	Test set	Train set	Test set	Train set	Test set	Train set	Test set
Ordering	20,625	2,805	1,009	161	17,546	1,910	39,402	4,175
Duration	456	75	21	3	374	51	729	106

Results

- Comparative experiments demonstrate that TACV outperforms existing state-of-the-art methods by a large margin.

	T-FEVER		T-FEVEROUS	
Methods	Label acc.	FEVER score	Label acc.	FEVER Score
KGAT	44.28	33.61	15.69	4.59
CGAT	44.38	33.91	16.58	4.29
ITR	44.05	30.88	31.66	8.63
UnifEE	49.67	41.10	49.14	**17.67**
TACV	**52.15**	**41.42**	**54.01**	15.38

- TACV remains robust on the original FEVER and FEVEROUS datasets.
- TACV shows superior performance on the real world LIAR and T-LIAR datasets, raising the confidence that TACV can be used for the verification of real world temporal claims.

	FEVER		FEVEROUS		LIAR	T-LIAR
Methods	Label acc.	FEVER score	Label acc.	FEVER score	Label acc.	Label acc.
KGAT	74.07	70.38	34.94	11.25	46.20	69.44
CGAT	76.39	73.15	39.70	12.52	45.77	72.22
ITR	73.36	70.04	44.20	14.39	49.24	69.44
TACV	**76.42**	**73.16**	**53.97**	**15.08**	**62.86**	**83.33**

Case Study

Claim: Illinois suffered 1,652 overdose deaths in 2014, of which 40 percent were associated with heroin and Illinois is ranked number one in the nation for a decline in treatment capacitybetween 2007 and 2012.
Ground Truth: SUPPORT

Method	Events	Retrieved Sentences	Event Label	Claim Label
TACV	• Illinois suffered 1,652 overdose deaths in 2014 , of which 40 percent were associated with heroin • Illinois ranked number one in the nation for a decline in treatment capacity between 2007 and 2012.	**•Illinois suffered 1,652 overdose deaths in 2014 – a 30 percent increase over 2010 – of which 40 percent were associated with heroin** •Durbin claims 40 percent of drug overdose deaths in Illinois involve heroin •However, the Illinois Department of Public Health, which reports preliminary and final drug overdose deaths to the CDC, puts the 2010 total at 1,284 and 1,700 in 2014 **•A report published in August 2015 by ICDP, shows state-funded treatment capacity in Illinois fell by 52 percent from 2007-2012, the largest decrease in the nation** •In 2007, Illinois ranked 28th in state-funded treatment capacity before dropping to No. 44, or third worst in 2012. •Durbin is correct when he says Illinois led the nation in the decline for state-funded treatment capacity.	SUP SUP	SUP
CGAT	-	**•Illinois suffered 1,652 overdose deaths in 2014 – a 30 percent increase over 2010 – of which 40 percent were associated with heroin** • The percent increase from 2010 is slightly more than 32 percent, and drug overdose deaths in 2014 that were associated with heroin is about 42 percent •In 2007, Illinois ranked 28th in state-funded treatment capacity before dropping to No. 44, or third worst in 2012	-	REF

- Claim has two events: "suffered in 2014" (in blue) and "ranked between 2007 and 2012" (in red).
- TACV is able to retrieve evidence sentences that confirm the date of overdose deaths for the first event, and sentences that mention the period when Illinois is ranked number one for decline in treatment capacity.
- The LLM verifies each event as SUPPORT, allowing TACV to correctly predict the overall claim label as SUPPORT.
- On the other hand, CGAT fails to retrieve sentences that reference the date when Illinois was ranked first for declined treatment capacity, leading to an incorrect prediction.

Conclusion

- We have proposed **TACV** for temporal fact verification that addresses the growing challenge posed by misinformation in real-world settings, particularly in information-heavy industries such as media, finance, and legal sectors.
- We have developed two temporal datasets that serve as evaluation benchmark for future research.
- Experimental results have demonstrated the effectiveness of **TACV** across multiple dataset including the real world Liar dataset.

Link to the project page