Estimating Temporal Validity of Text

Adam Jatowt

Univ. of Innsbruck

8 March 2024

EXPIRY DATE

Few months unless kept updated.

Today’s Agenda

1. Introduction & Background
   - Temporal Commonsense Reasoning of LLMs
2. Novel tasks
   1. Temporal text validity
   2. Temporal text validity reassessment
   3. Temporal validity change prediction
3. Injecting time into LLMs
4. Conclusions

Time & Text

- Time is of the essence:
  - A key aspect of stories, events, narrative documents, message streams, etc.
  - Can determine if documents are relevant
  - Can tell how the different story pieces should be combined
  - Can help in correct text understanding & and information extraction…
  - …

So far, the NLP and IR communities placed rather limited focus on research towards understanding and utilizing temporal aspects of text…

Two Main Types of Temporal Knowledge

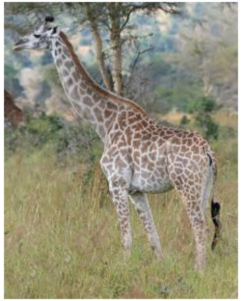
- Temporal Commonsense Knowledge
  - E.g., visiting a doctor is <u>after</u> breaking the leg rather than before
  - E.g., going on holiday takes <u>longer than</u> going for a walk
- Temporal Factual Knowledge
  - E.g., Barack Obama was the US president <u>from 2009 to 2017</u>
  - E.g., Hiroshima and Nagasaki bombings were <u>after</u> the Attack on Pearl Harbor

Commonsense Reasoning

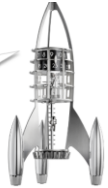
- the basic level of practical knowledge and reasoning
- concerning common situations and events
- that are commonly shared among most people

For example, it's ok to keep the closet door open,
but it's not ok to keep the fridge door open,
as the food inside might go bad.

<



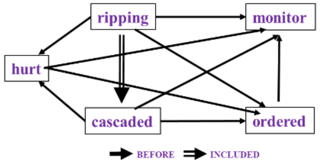
You don't reach to the moon
by making the tallest building in the world
taller

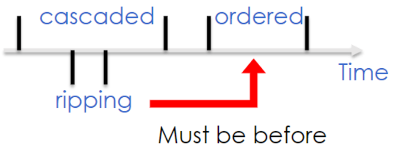


Y. Choi et al., ACL 2020 Commonsense Tutorial, https://maartensap.com/acl2020-commonsense/

Temporal Commonsense Reasoning Examples

In Los Angeles that lesson was brought home today when tons of earth **cascaded** down a hillside, **ripping** two houses from their foundations. No one was **hurt**, but firefighters **ordered** the evacuation of nearby homes and said they'll **monitor** the shifting ground until March 23rd.





https://maartensap.com/acl2020-commonsense

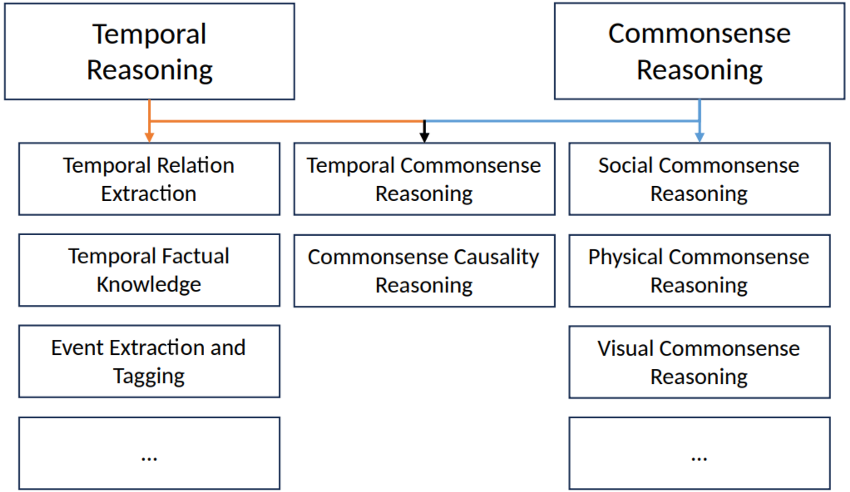
Dr. Porter is **taking a vacation** and _______ be able to see you soon.

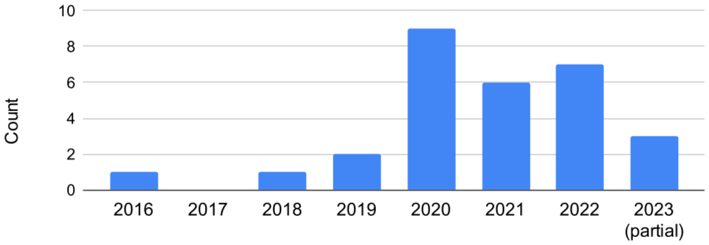
Dr. Porter is **taking a walk** and ___ be able to see you soon.

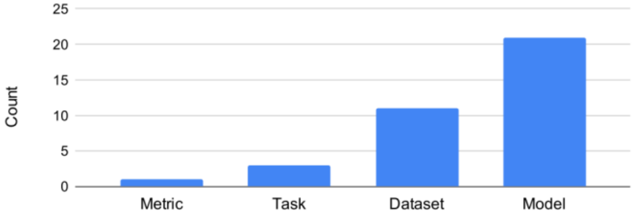
https://maartensap.com/acl2020-commonsense

Y. Choi et al., ACL 2020 Commonsense Tutorial, https://maartensap.com/acl2020-commonsense/
G. Wenzel, A. Jatowt: “An Overview of Temporal Commonsense Reasoning and Acquisition”, 2023, arxiv

Temporal Commonsense Reasoning Research Area







G. Wenzel, A. Jatowt: "An Overview of Temporal Commonsense Reasoning and Acquisition", 2023, arxiv

Temporal Commonsense Reasoning

A language model with a robust understanding of temporal context is primed to perform better on downstream NLP tasks such as:

- storytelling (Mostafazadeh et al., 2016)
- natural language inference (Hosokawa et al., 2023)
- timeline understanding (Steen and Markert, 2019)
- user status tracking (Xia and Qi, 2022)
- dialogue management
- etc.

Types of Commonsense Reasoning

- Different types of temporal commonsense reasoning tasks (Zhou et al, 2019):
  - **Event Duration** (ED):
    - reasoning about event durations.
  - **Event Ordering** (EO):
    - reasoning about the typical sequence of events.
  - **Frequency** (F):
    - reasoning about the frequency of event occurrences.
  - **Stationarity** (S):
    - reasoning about the length of state persistence.
  - **Typical Time** (TT):
    - reasoning about the specific timing of events.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. "Going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. EMNLP-IJCNLP 2019

Temporal Text Validity

Temporal Validity Estimation: News Example

Assume you read the following sentences in a newspaper published 1 month ago:

- *“Chancellor of Austria is visiting France."*
- *“France is a member of United Nations.”*

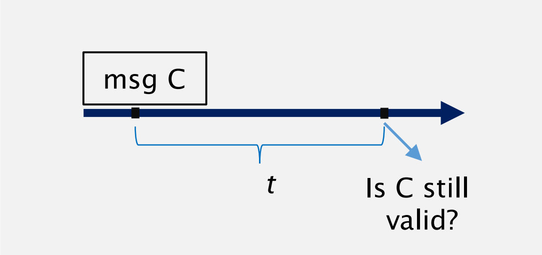
Which would you consider still as true?

Tweet Example

Assume you have just received the following message from your friend:

– *“I am taking a walk."*

How long would you consider it as still being true?



Temporal Validity Definition

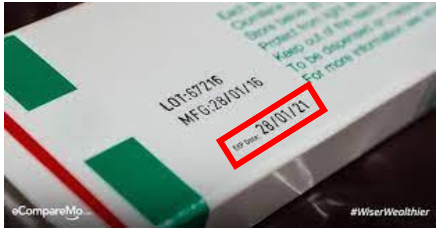
Temporal validity is a measure of how long the information remains valid after it has been created or expressed

Given a content C created at time t, its validity period is the **<u>maximum time</u>** after t during which the information expressed in C **<u>remains valid</u>**

Temporal Validity of Text as Information Expiry Date

Analogy to Product's Expiry Date:

Important concept determining the time until which a product, or more generally an object, remains usable



Axel Almquist, Adam Jatowt: *Towards Content Expiry Date Determination: Predicting Validity Periods of Sentences*. ECIR 2019, pp. 86-101

Example Temporal Validity Applications

- Measuring text obsoleteness
- Recommender systems
- Enhancing information retrieval (e.g., filtering Twitter timeline)
- Conversational AI
- Fact checking
- Story understanding
- Etc.

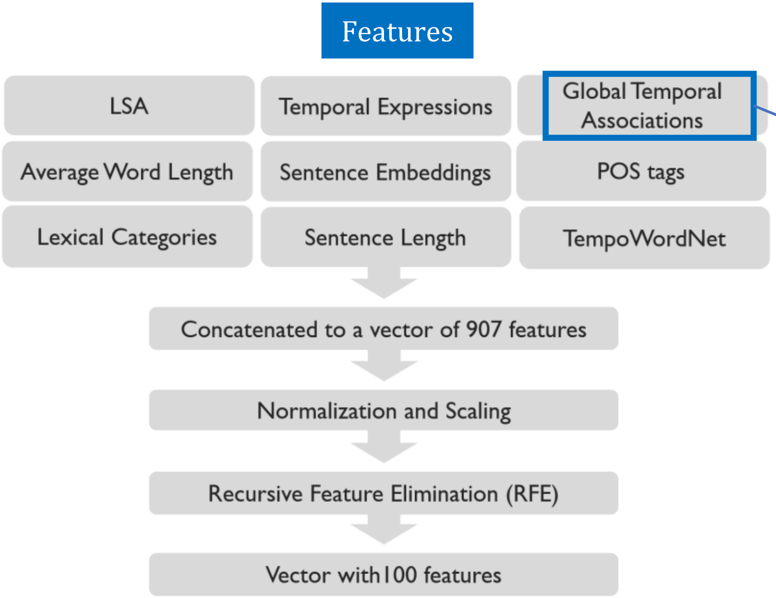
Validity Classes

Validity Classes

Within a few hours	*"So Michi, Audrey, Joel and myself are all hanging out in Linda's basement."*
Within a few days	*"School starts at a later time on Wednesday but that's no big deal."*
Within a few weeks	*"I am taking a course on learning how to use the program 3d studio max."*
Within a few months	*"I am also playing a gig with the new millennium string orchestra at the beginning of next month."*
Within a few years or more	*"The middle eastern nation of Israel is planning to expand its settlements, its housing areas in the west bank."*

Axel Almquist, Adam Jatowt: *Towards Content Expiry Date Determination: Predicting Validity Periods of Sentences*. ECIR 2019, pp. 86-101

Simple Approach for Estimating Temporal Validity



Computed by measuring **co-occurrences of duration temporal expressions** and **words** in Common Crawl

Table 2: Statistics about sentences extracted from Common Crawl.

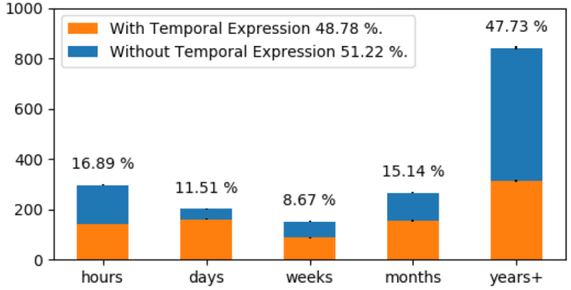
Sentences	Sentences with temp exp	Temp exp	Temp exp as verb modifiers
8,800,000	1,811,608	2,249,309	823,944

Axel Almquist, Adam Jatowt: *Towards Content Expiry Date Determination: Predicting Validity Periods of Sentences*. ECIR 2019, pp. 86-101

Temporal Validity Estimation Datasets

Estimate how long an action expressed in a sentence would typically take place

- Task: classification
- Classes:
  - a) hours, days, weeks, months, years (or longer)
  - b)
- Size:
  - a) 1.7k
  - b)
- Source:
  - a) blogs, news, Wikipedia
  - b)
- Generation way: crowdsourcing



Almquist and Jatowt: *Towards Content Expiry Date Determination: Predicting Validity Periods of Sentences*. ECIR 2019
Lynden et al.: *Commonsense Temporal Action Knowledge (CoTAK) Dataset*, CIKM 2023

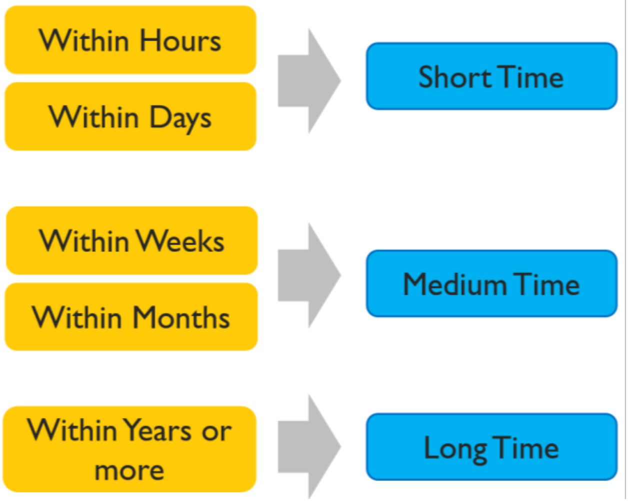
Experimental Results

Result using original classes

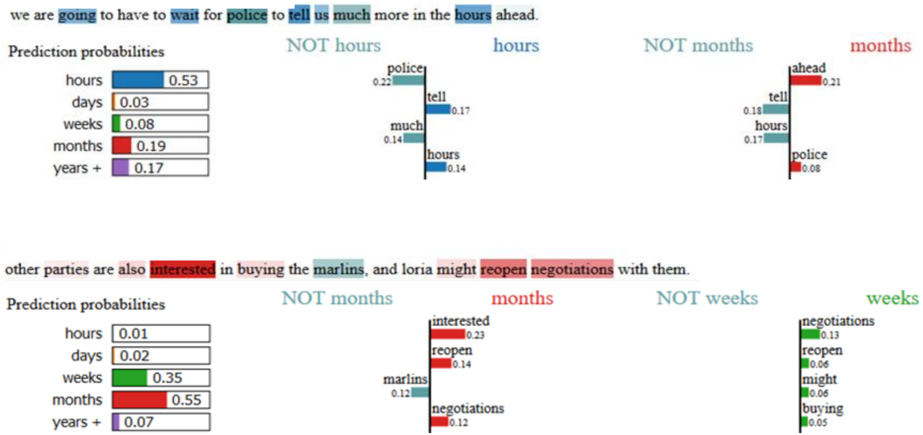
Models	F1-micro
Random	19.61
Majority Class	47.76
RNN	59.49
MLP (LSA)	39.17**
KNN (LSA)	56.95**
RandomForest (LSA)	60.01**
SVC_RBF (LSA)	61.77**
LinearSVC (LSA)	62.39**
MLP (all features)	53.76**
KNN (all features)	60.07**
RandomForest (all features)	62.75**
SVC_RBF (all features)	67.44**
LinearSVC (all features)	68.69**

Result using reduced classes

Models	F1-micro
Random	34.94
Majority Class	50.14
RNN	70.51
MLP (LSA)	63.61**
KNN (LSA)	61.77**
RandomForest (LSA)	66.15**
SVC_RBF (LSA)	69.48**
LinearSVC (LSA)	70.11**
MLP (all features)	72.50**
KNN (all features)	68.75**
RandomForest (all features)	70.90**
SVC_RBF (all features)	77.37**
LinearSVC (all features)	78.11**



Examples of Classified Sentences



LIME: https://github.com/marcotcr/lime

Temporal Validity Estimation Datasets

Estimate how long an action expressed in a sentence would typically take place

- Task: classification
- Classes:
  a) hours, days, weeks, months, years (or longer)
  b) seconds, minutes, hours, days, longer
- Size:
  a) 1.7k
  b) >300k
- Source:
  a) blogs, news, Wikipedia
  b) WikiHow sentences
- Generation way: crowdsourcing

Examples: “lift the foot”, “remove the old shoe”, “clean the hoof”, etc.

Table 1: Fine-tuned BERT prediction accuracy

Serialization	Coarse-grained		Fine-grained	
	perform	effect	seconds	minutes
Action desc. only	0.79	0.76	0.51	0.81
Title + description	0.83	0.81	0.55	0.81

Almquist and Jatowt: *Towards Content Expiry Date Determination: Predicting Validity Periods of Sentences*. ECIR 2019
Lynden et al.: *Commonsense Temporal Action Knowledge (CoTAK) Dataset*, CIKM 2023

Current Work: Multimodal task extension

- Building a dataset

<

Expected time needed to do an action: “tidy the room”

Current Work: Multimodal task extension

- Building a dataset

>

Expected time needed to do an action: “cook a cake”

Current Work: Obsoleteness Prediction in Text

Input text

Brest is a port city in the Finistère department, Brittany. Located in a sheltered bay not far from the western tip of a peninsula and the western extremity of metropolitan France, Brest is an important harbour and the second largest French military port after Toulon. The city is located on the western edge of continental France. With 139,456 inhabitants (2020), Brest forms Western Brittany's largest metropolitan area (with a population of 370,000 in total), ranking third behind only Nantes and Rennes in the whole of historic Brittany, and the 25th most populous city in France (2019); moreover, Brest provides services to the one million inhabitants of Western Brittany.
François Cuillandre is the mayor of the city..

Current Work: Obsoleteness Prediction in Text

Input text

Brest is a port city in the Finistère department, Brittany. Located in a sheltered bay not far from the western tip of a peninsula and the western extremity of metropolitan France, Brest is an important harbour and the second largest French military port after Toulon. The city is located on the western edge of continental France. With 139,456 inhabitants **(**2020**),** Brest forms Western Brittany's largest metropolitan area (with a population of 370,000 in total), ranking third behind only Nantes and Rennes in the whole of historic Brittany, and the **25th most populous city in France** **(**2019); moreover, Brest provides services to the one million inhabitants of Western Brittany.
François Cuillandre is the mayor of the city..

Temporal Text Validity Reassessment

Temporal Text Validity Reassessment

Assume you have just received the following message from your friend:

- Post A: "*I am taking a walk* “
- Post B: “*Getting a cup of coffee for take-away*”
- Post C: “*Just started preparing dinner*”

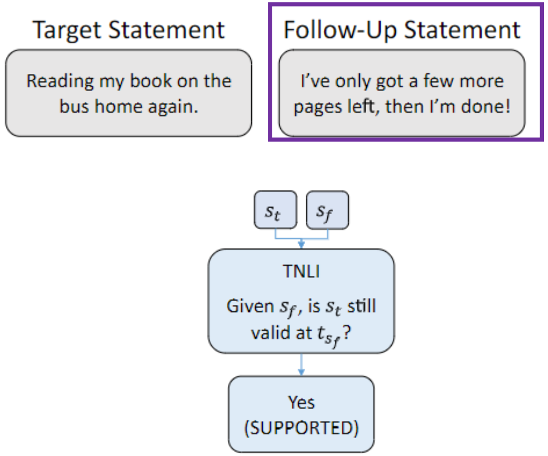
Is the post A still true after reading post B?

How about after reading post C?

Context-based Temporal Validity Prediction: Temporal Text Validity Reassessment

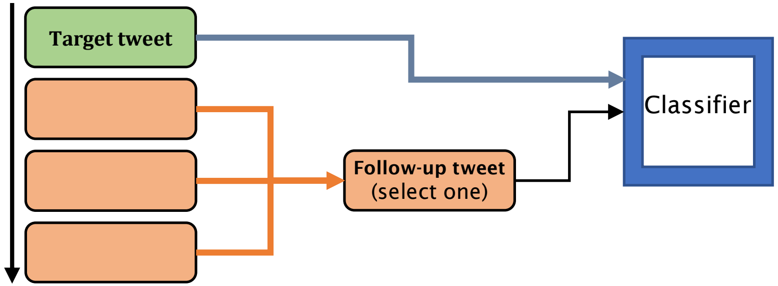
Estimate if an action expressed in a sentence would continue or cease to take place in view of additional context

- Task: classification
- Classes: `supported`, `invalidated`, `neutral`



Taishi Hosokawa, Adam Jatowt, Kazunari Sugiyama: *Temporal Natural Language Inference: Evidence-based Evaluation of Temporal Text Validity*, ECIR 2023

Example Application Scenario



Does any of the follow-up tweet **render** the **original tweet obsolete**?

Similarity to Natural Language Inference (aka. Text Entailment)

Natural Language Inference (NLI): *task of determining the inference relation between two short texts*
3 classes: `Entailment`, `Contradiction` and `Neutral`

Premise	Hypothesis	Label
A soccer game with multiple males playing.	Some men are playing a sport.	ENTAILMENT
A black race car starts up in front of a crowd of people.	A man is driving down a lonely road.	CONTRADICTION
A smiling costumed woman is holding an umbrella.	A happy woman in a fairy costume holds an umbrella.	NEUTRAL

Hypothesis	Premise	Label
A small Asian street band plays in a city park.	Their performance pulls a large crowd as they used some new tunes and songs today.	SUPPORTED
A woman in blue rain boots is eating a sandwich outside.	She takes off her boots in her house.	INVALIDATED
A man jumping a rail on his skateboard.	His favorite food is pizza.	UNKNOWN

https://nlp.stanford.edu/projects/snli/

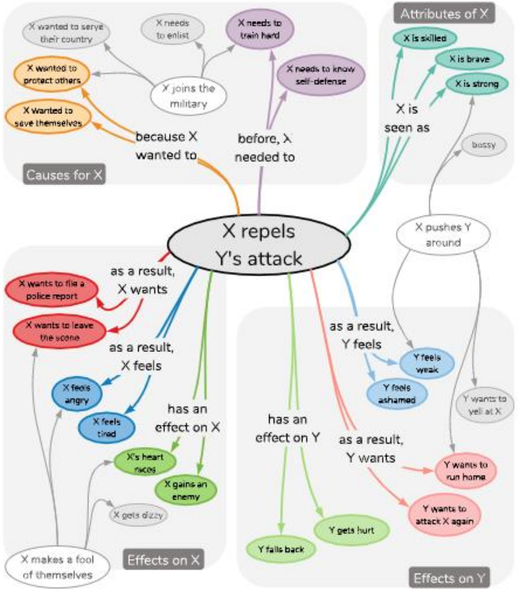
ATOMIC Knowledge Base

Commonsense Reasoning Knowledge Base (1.33M commonsense knowledge tuples and 23 relations)

- Includes ConceptNet knowledge
  - ConceptNet - the most commonly used commonsense knowledge base about physical entities

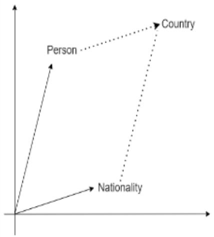
- 9 if-then relation classes
  - cause
  - reaction
  - intention
  - effect
  - …

- Many physical and event-centered relations
  - “IsAfter”, “IsBefore”, “HasSubEvent”, “HinderedBy”



https://homes.cs.washington.edu/~msap/atomic/

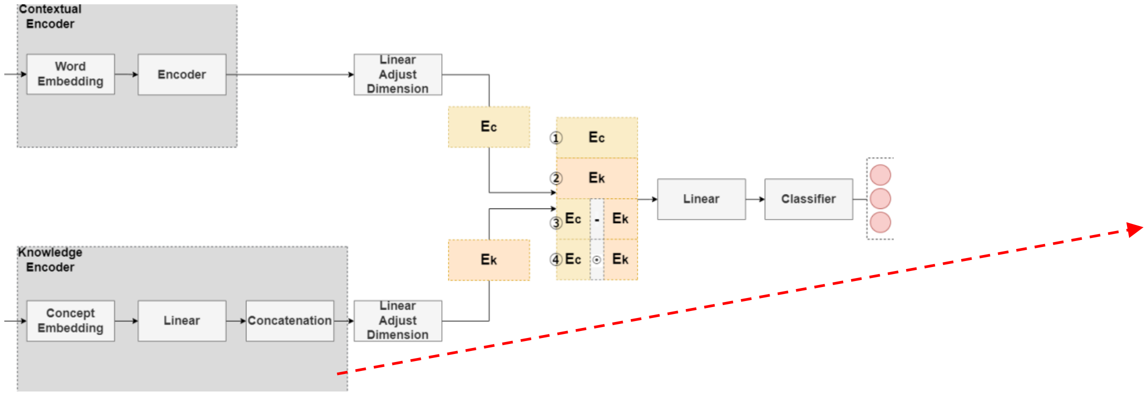
Proposed Model



Embedding in knowledge bases (TransE):
head entity + relation = tail entity

Learns information from a knowledge base
- Uses embeddings for representing data in knowledge base
- Knowledge base: tuples <head entity, relation, tail entity>

Combines text-based method (e.g., BERT) with knowledge-based encoding method (e.g., TransE)



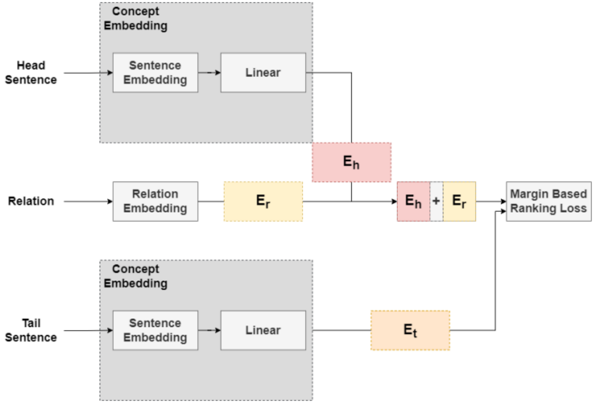


Fig. 2: TransE model for sentences.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, *26*.

Initial Analysis using NLI-based Pretraining

- NLI-based pretraining improves Siamese network’s performance
- NLI-based pretraining does not improve the result of Self-Explaining model (SOTA model for NLI)
    - NLI-related information may be already included during the pre-training for Self-Explaining model

Model	Accuracy
Siamese	0.715
+SNLI	0.756
+MNLI	0.757
Self-Explaining	0.873
+SNLI	0.867
+MNLI	0.535

SUPPORTED = Entailment (NLI)
INVALIDATED = Contradiction (NLI)
UNKNOWN = Neutral (NLI)

Sun, Zijun, et al. "Self-explaining structures improve nlp models." *arXiv preprint arXiv:2012.01786* (2020).

Constructing Dataset

- Hypotheses: randomly sampled 5k premises from SNLI
- Premises: created through crowdsourcing with AMT for each of 3 classes and subject to manual verification
- Result: **10,659** sentence pairs balanced over 3 classes

Hypothesis	Premise	Label
A small Asian street band plays in a city park.	Their performance pulls a large crowd as they used some new tunes and songs today.	**SUPPORTED**
A woman in blue rain boots is eating a sandwich outside.	She takes off her boots in her house.	**INVALIDATED**
A man jumping a rail on his skateboard.	His favorite food is pizza.	**UNKNOWN**

Table 4.1: The length of sentences.

	Average	Variance
hypothesis	11.4	19.4
premise	8.9	10.8
invalidated	8.4	8.6
supported	9.3	10.7
unclear	8.9	12.7

Experimental Results

- Self-explaining model performs best
- Siamese + TransE performs better than Siamese
- Self-explaining + TransE has almost same accuracy as self-explaining

Ground truth

Predicted

	Inv^G	Sup^G	Unk^G
Inv^P	2866	580	107
Sup^P	857	2057	642
Unk^P	270	580	2703

Siamese

	Inv^G	Sup^G	Unk^G
	2824	585	144
	564	2641	348
	259	414	2880

Siamese + TransE

Model	Pre-Train Loss	Accuracy
TransE	0.19	0.878
TransH	0.48	0.868
ComplEx	1.24	0.856

Tested variants of TransE
(in conjunction with Self-Explaining model)

Model	Accuracy
Siamese	0.715
SBERT + FFN	0.806
BERT	0.441
Self-Explaining BERT	0.805
Self-Explaining RoBERTa	0.873
Siamese+TransE	0.784
Self-Explaining BERT+TransE	0.819
Self-Explaining RoBERTa+TransE	0.878
GPT 3.5	0.620
Llama	0.320

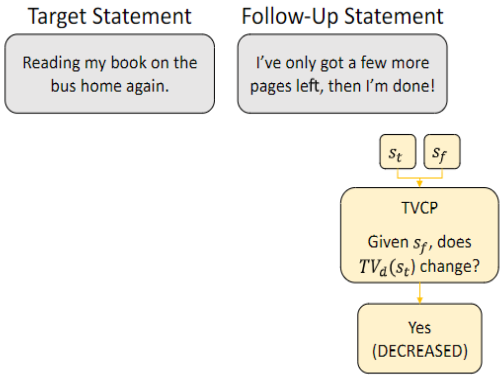
Incorporating TransE to Siamese net helps to more correctly determine SUPPORTED and UNKNOWN classes (improvement of 28% and 6.5%), while only slightly confusing the INVALIDATED class (decrease of 1.4%)

Temporal Text Validity Change Prediction

Temporal Validity Change Prediction

Estimate if an action expressed in a sentence would **increase** or **decrease** the temporal validity of another sentence

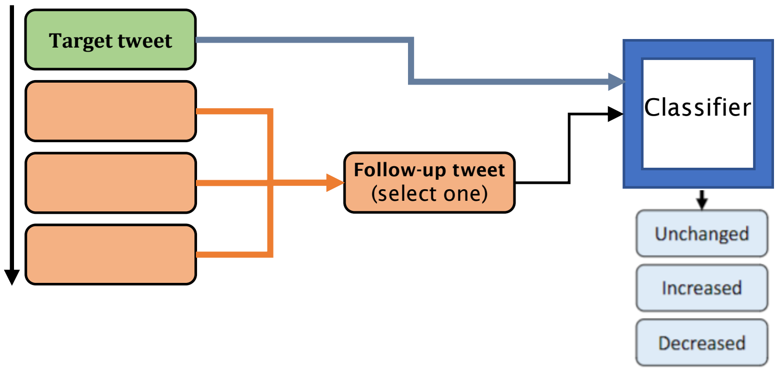
- Task: classification
- Classes: `decreased`, `increased`, `neutral`



$$\mathrm{TVCP}(s_t, s_f) = \begin{cases} \mathrm{DEC} & \mathrm{TV}_d(s_t) > \mathrm{TV}_d^{s_f}(s_t) \\ \mathrm{UNC} & \mathrm{TV}_d(s_t) = \mathrm{TV}_d^{s_f}(s_t) \\ \mathrm{INC} & \mathrm{TV}_d(s_t) < \mathrm{TV}_d^{s_f}(s_t) \end{cases}$$

Georg Wenzel, Adam Jatowt: Temporal Validity Change Prediction. CoRR abs/2401.00779 (2024) https://arxiv.org/pdf/2401.00779.pdf

Application Scenario



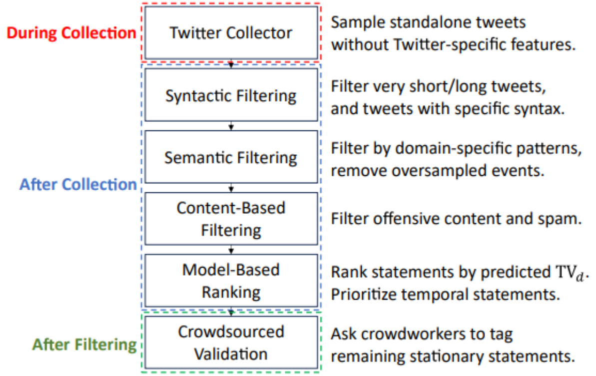
Does any of the follow-up tweet **changes** the temporal validity duration of the **original tweet**?

"I have a doctors appt in 10 minutes which means I might have to wait for an hour"

- ✚ **"Tick tock. This is always so boring, no one ever tells me how long I am going to be here."** No change in expected validity.
- ✚ **"The doctor was actually on time for once, so glad the visit is going quickly"** **Decreased** expected validity
- ✚ **"The doctor got called away on an emergency. Will have to reschedule later this week."** **Increased** expected validity

Dataset building

- Size: 5k
- Source: sentences from Twitter
- Generation way: crowdsourcing



$$t \in \{< 1 \text{ minute}, \text{1-5 minutes}, \text{5-15 minutes}, \text{15-45 minutes}, \text{45 minutes-2 hours}, \text{2-6 hours}, \text{more than 6 hours}, \text{1-3 days}, \text{3-7 days}, \text{1-4 weeks}, \text{more than 1 month}\}$$

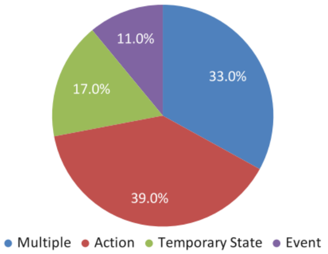
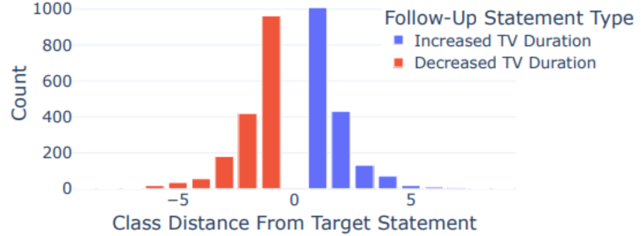


Figure 2: Distribution of different types of temporal information in a sample of our dataset



Results

Model	$\overline{\text{Acc}}$ (+ MT)	$\overline{\text{EM}}$ (+ MT)
TF - RoBERTa	64.0 (+1.5)	21.2 (+2.5)
ChatGPT	66.3 (N/A)	29.3 (N/A)
S - RoBERTa	78.7 (+1.1)	48.2 (+2.1)
TF - CoTAK	83.2 (+0.6)	58.2 (+1.4)
S - BERT	83.8 (−0.3)	59.1 (−1.5)
TF - TacoLM	83.5 (+1.4)	59.1 (+2.9)
TF - BERT	84.8 (−0.2)	61.2 (+0.9)
SelfExplain	88.5 (+1.1)	69.8 (+2.8)

Table 3: Model evaluation results, sorted by mean EM score. TF = TransformerClassifier, S = SiameseClassifier, MT = Multitask Implementation

Figure 8: Training data vs. performance metrics in MultiTask

Figure 9: Temporal validity change delta vs. accuracy in MultiTask, SelfExplain and ChatGPT

COINTELEGRAPH
The future of money

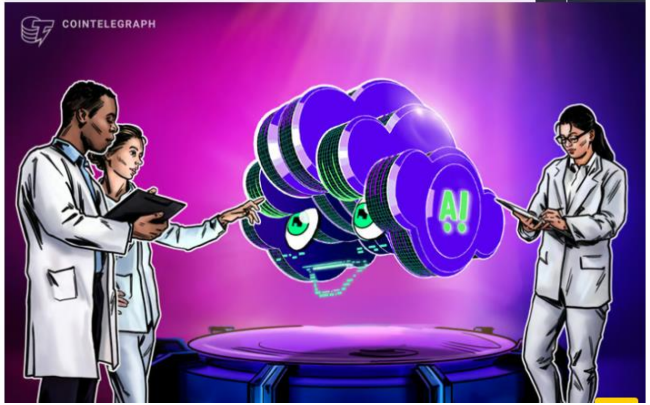
BTC	ETH	XRP	BNB	AD
$43,919	$2,248	$0.57	$307	$0.
+0.18%	+0.01%	+0.27%	-1.92%	-1.8

News · Markets · Magazine · People · Cryptopedia · Research · Video · Po

TRISTAN GREENE JAN 02, 2024

AI experiment involving 'temporal validity' could have significant implications for fintech

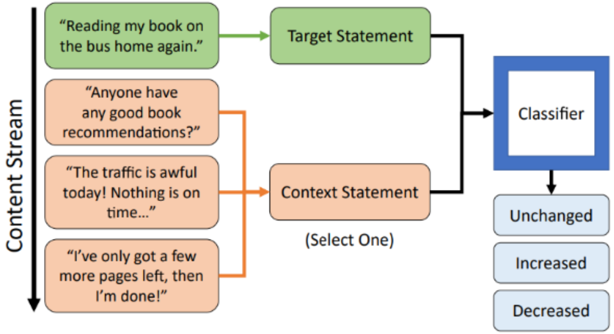
Teaching AI to understand the importance of timeliness could lead to better prediction models.



A pair of researchers from the University of Innsbruck in Austria have developed a method to determine how well an artificial intelligence (AI) system is at understanding 'temporal validity,' a benchmark that could have significant implications for the use of generative AI products such as ChatGPT in the fintech sector.

Temporal validity refers to how relevant a given statement is to another statement over time. Essentially, it refers to the time-based value of paired statements. An AI being evaluated on its ability to predict temporal validity would be given a set of statements and asked to choose the one most closely related through time.

In their recently published pre-print research paper titled "Temporal Validity Change Prediction," Georg Wenzel and Adam Jatowt use the example of a statement wherein a person is declared to be reading a book on a bus.



In the above example, the most valid context statement is "I've only got a few more pages left, then I'm done." As the target statement indicates the bus rider is currently reading a book, the other two are irrelevant by comparison. Image source: Wenzel, Jatowt 2024.

https://cointelegraph.com/news/ai-artificial-intelligence-experiment-involving-temporal-validity-significant-implications-fintech

Incorporating Time into LLMs

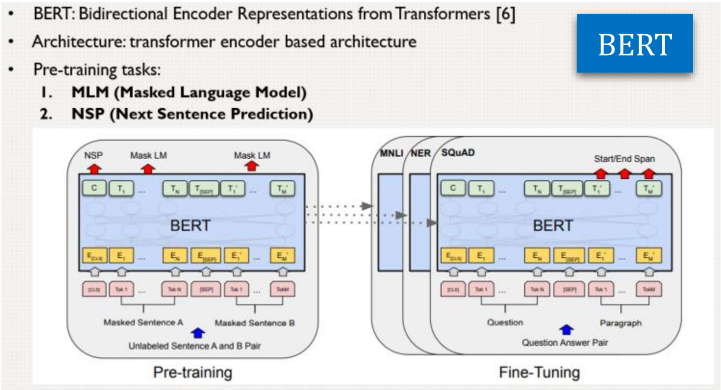
Time and LLMs

- How to design LLMs that would pay more attention on temporal aspects in text?

- Pretraining corpora?
- Pretraining tasks?
- Attention mechanism change?
- ...

Domain-specific LLMs

- Some language models use specialized pre-training tasks for particular tasks\domains:
  - SpanBERT: replaces Masked Language Model (MLM) with Span Masking to obtain SOTA performance on span selection tasks such as machine reading comprehension
  - SentiLARE: extends MLM to label-aware MLM, and obtains new SOTA performance on sentiment analysis tasks



Existing language models do not seem to explicitly utilize temporal aspects of text

BiTimeBERT

- Architecture: transformer encoder
- Objective: enhance language representations with temporal information
- Pre-training dataset: NYT Corpus (1.8 million news)

Novel pre-training tasks:

1. Time-aware Masked Language Modeling (TAMLM)

Explicitly introduces temporal expressions in text during pre-training: mask 30% of temporal expressions in text

2. Document Dating (DD)

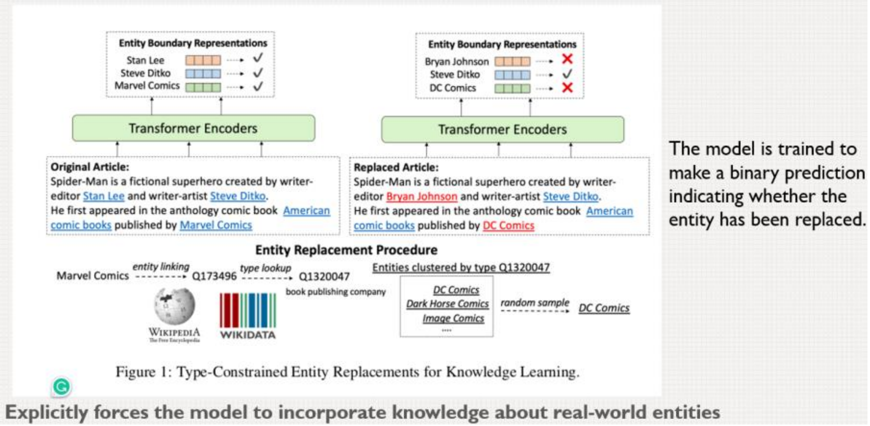
Incorporates document timestamp information during pre-training: predict timestamp at given granularity

J. Wang, A. Jatowt, M. Yoshikawa, Y. Cai: *BiTimeBERT: Extending Pre-Trained Language Representations with Dual-Type Temporal Information*. SIGIR 2023

TIR Task (not included in the final model)

3. TIR: Temporal Information Replacement

- 50% of temporal expressions are replaced by other temporal expressions of the same granularity.
- The task is to predict whether the temporal expressions are replaced.



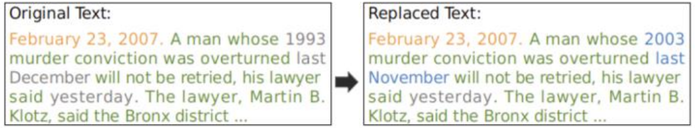


Figure 2: Example of the replacement procedure in TIR task.

WLKM: Weakly Supervised Knowledge-Pretrained Language Model [1]

[1] Xiong, Wenhan, et al. *Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model*. arXiv preprint arXiv:1912.09637 (2019).

Datasets

- 5 temporal datasets of different character
  - Event time estimation
  - Document timestamping (dating)
  - Temporal question answering

Table 2: Statistics of the datasets.

Dataset	Size	Time Span	Source	Granularity	Task
EventTime	22,398	1987-2007	Wikipedia & "On This Day" Website	Day, Month, Year	Event Time Estimation
WOTD	6,809	1302-2018	Wikipedia Website	Year	Event Time Estimation
NYT-Timestamp	50,000	1987-2007	News Archive	Day, Month, Year	Document Dating
TDA-Timestamp	50,000	1785-2009	News Archive	Day, Month, Year	Document Dating
NYT-Corpus	*1.8 Million*	*1987-2007*	*News Archive*	*Day, Month, Year*	*Pre-training*

Performance of BiTimeBERT

Table 3: Performance of different models on EventTime datasets with two different settings.

Model	EventTime				EventTime-WithTop1Doc			
	Year		Month		Year		Month	
	ACC	MAE	ACC	MAE	ACC	MAE	ACC	MAE
RG	4.77	6.92	0.41	81.60	4.77	6.92	0.40	81.70
BERT	21.65	3.47	5.09	43.81	35.98	3.89	5.98	37.95
BERT-NYT	21.25	3.56	5.18	43.50	34.46	4.45	8.21	34.14
SOTA [54]	-	-	-	-	40.93	3.01	**30.89**	36.19
BERT-TIR	25.40	3.23	6.83	40.45	36.47	3.54	17.01	31.72
BiTimeBERT	**31.91**	**3.12**	**12.99**	**34.79**	**41.96**	**2.40**	25.76	**28.86**

Event dating

Table 5: Performance of different models for document dating on NYT-Timestamp and TDA-Timestamp.

Model	NYT-Timestamp				TDA-Timestamp			
	Year		Month		Year		Month	
	ACC	MAE	ACC	MAE	ACC	MAE	ACC	MAE
RG	4.77	7.06	0.41	81.79	0.45	75.39	0.04	873.88
BERT	35.00	1.64	2.56	22.74	15.84	44.87	0.80	632.66
BERT-NYT	38.74	1.41	8.24	18.35	15.04	45.16	0.66	669.02
BERT-TIR	48.06	1.09	20.30	13.54	17.72	43.53	1.26	589.69
BiTimeBERT	**58.72**	**0.80**	**31.10**	**9.54**	**19.00**	**40.11**	**2.38**	**580.25**

Document timestamping

Table 14: Sentence time prediction results.

Model	NYT-years		
	1981-2020	1987-2007	1981-1986 & 2008-2020
	ACC	ACC	ACC
BERT	10.02	9.7	10.38
BERT-NYT	10.23	10.75	9.64
TempoBERT [41]	9.24	-	-
BiTimeBERT	**12.52**	**13.44**	**11.51**

Sentence dating

Model	Top 1		Top 5		Top 10		Top 15	
	EM	F1	EM	F1	EM	F1	EM	F1
QANA [53]	21.00	28.90	28.20	36.85	34.20	44.01	36.20	45.63
QANA+BiTimeBERT	**22.40**	**29.31**	**29.20**	**37.14**	**34.80**	**44.34**	**36.40**	**46.01**

Temporal question answering

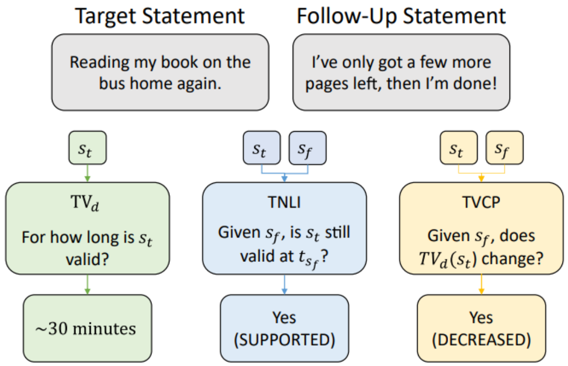
J. Wang, A. Jatowt, M. Yoshikawa, Y. Cai: *BiTimeBERT: Extending Pre-Trained Language Representations with Dual-Type Temporal Information*. SIGIR 2023

Conclusions

Temporal reasoning tasks related to temporal commonsense reasoning in text:

1. Temporal validity duration prediction
2. Temporal validity reassessment
3. Temporal validity change prediction

Incorporating time into LLMs



Future work:

1. Creation of large-scale, complex datasets
2. Extension to reason about future
3. Prompt engineering techniques for improving temporal reasoning of LLMs