SECTION 5 · 25 min

Neural & Transformer-based Temporal Models

How transformer encoders learn — and forget — temporal signals through conditioning, masking, and adaptation.

Presenter: Avishek Anand

05

Pre-Trained LMs

THE TRANSITION

What changed with the coming of contextual models?

We shifted from **count-based statistical LMs** to powerful **contextual models**. While language understanding drastically improved, models were pre-trained on massive static corpora without temporal structures, creating severe **temporal blindspots**.

The Core Research Question: How to inject temporal dimensions?

01

Pre-Training

Integrating temporal signals directly into the primary training phase (e.g., parameterizing objectives like MLM with timestamps).

02

Post-Training

Adapting and updating models post-pre-training to align with fresh temporal context, preventing knowledge stagnation.

03

Retrieval or Inductive Biases

Designing model architectures or structuring training data to inherently respect and represent temporal dynamics and order.

Time-aware Ranking

LMs are trained on a **snapshot** of the web — but many facts come with an expiration date.

"Cristiano Ronaldo plays for ___ "

2012 → Real Madrid 2019 → Juventus

Averaging
Conflicting time-scoped facts collapse into one low-confidence answer.

Forgetting
Skewed corpora (more recent docs) → facts from the distant past are lost.

Poor calibration
As models go stale, they are queried about facts outside training scope.

Temporal drift
Knowledge is volatile; static corpora cannot track evolving truth.

THE SHIFT

From treating time as an **auxiliary signal** → to making it a **core component of the model**.

Three levers: (1) condition on timestamps · (2) re-engineer objectives & attention · (3) ground generation in time.

Time-aware Language Models

All TLMs inject time during pretraining or fine-tuning — they differ in WHERE the temporal signal enters.

1 · Timestamp Conditioning

Parametrize P(y | x, t): feed the time alongside the text.

KEY MODELS

- **TempoT5 (Dhingra ’22)** ★
- **TempoBERT (Rosin ’22)**
- **BiTimeBERT (Wang ’23)** ★

2 · Architecture & Objectives

New masking, attention & representations that internalize temporal structure.

KEY MODELS

- **Temporal Attention (Rosin ’22)**
- **Temporal Span Masking (Cole ’23)**
- **TALM (Ren ’23) · SG-TLM (Su ’23)**

3 · Temporal Grounding for Generation

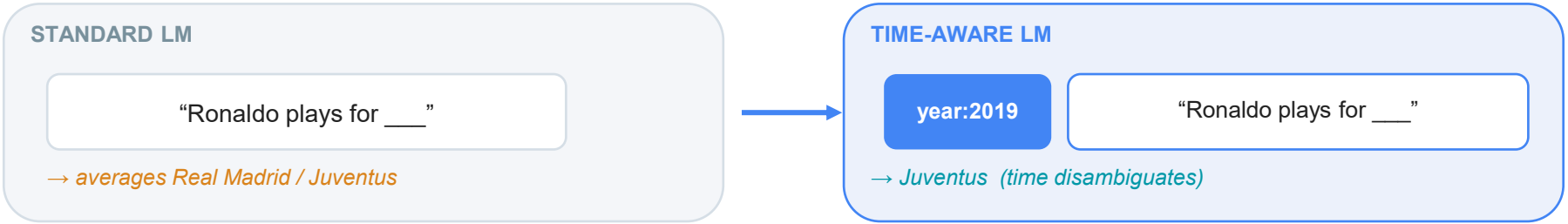
Steer generation with time-specific context.

KEY MODELS

- **Temporal Prompts (Cao & Wang ’22)**
- **TCQA (Son & Oh ’23)**

Timestamp Conditioning

Core idea: **reparametrize** **P(y | x) → P(y | x, t)** — let the model see the time, not just the text.



TWO WAYS TO INJECT THE TIMESTAMP

As input prefix / tokens

Prepend the date as text (**“year: 2019”**) or special tokens to the sequence.

e.g. TempoT5 · TempoBERT

As a pretraining task

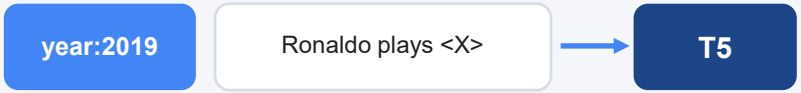
Make the model predict / use time via new objectives over a temporal corpus.

e.g. BiTimeBERT (TAMLM + Doc-Dating)

TempoT5

METHOD

Prefix the input with its time before T5 span corruption:



$P(y \mid x, t; \theta)$ y = salient span

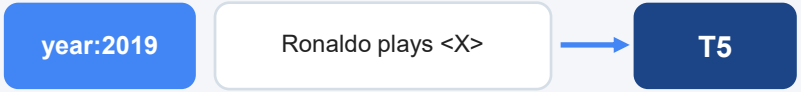
- Trained on CustomNews (2010–18) + synthetic TempLAMA probes
- Salient-span masking over named entities & dates
- Cheap “refresh” to new years — no full retraining

Contribution: **TempLAMA** — fill-in-the-blank probe whose answer changes over time

TempoT5

METHOD

Prefix the input with its time before T5 span corruption:



$P(y \mid x, t; \theta)$ y = salient span

- Trained on CustomNews (2010–18) + synthetic TempLAMA probes
- Salient-span masking over named entities & dates
- Cheap “refresh” to new years — no full retraining

Contribution: **TempLAMA** — fill-in-the-blank probe whose answer changes over time

WHAT IT FIXES

Memorization
Better recall of facts from the training period vs. a uniform model.

Calibration
Sensible confidence on unseen future facts (future ≈ present).

Updatability
Ingests new data without catastrophic forgetting.

TAKEAWAY

The simplest lever — a time prefix — already buys disambiguation, calibration & refreshability.

Bi-TimeBERT

Wang, Jatowt, Yoshikawa & Cai, SIGIR 2023 · pretrained on a 20-year news archive

KEY INSIGHT — documents carry TWO distinct temporal signals

Publication time
When the document was written (timestamp / dateline)

Content time
When described events happened (temporal expressions)

Bi-TimeBERT

Wang, Jatowt, Yoshikawa & Cai, SIGIR 2023 · pretrained on a 20-year news archive

KEY INSIGHT — documents carry TWO distinct temporal signals

Publication time
When the document was written (timestamp / dateline)

Content time
When described events happened (temporal expressions)

TWO NEW PRETRAINING OBJECTIVES

TAMLM — **Time-Aware Masked LM**

Mask temporal expressions and reconstruct them → bind language to content time.

DD — **Document Dating**

Predict the doc's publication timestamp → task-level temporal supervision.

Bi-TimeBERT

Wang, Jatowt, Yoshikawa & Cai, SIGIR 2023 · pretrained on a 20-year news archive

KEY INSIGHT — documents carry TWO distinct temporal signals

Publication time
When the document was written (timestamp / dateline)

Content time
When described events happened (temporal expressions)

TWO NEW PRETRAINING OBJECTIVES

TAMLM — **Time-Aware Masked LM**

Mask temporal expressions and reconstruct them → bind language to content time.

DD — **Document Dating**

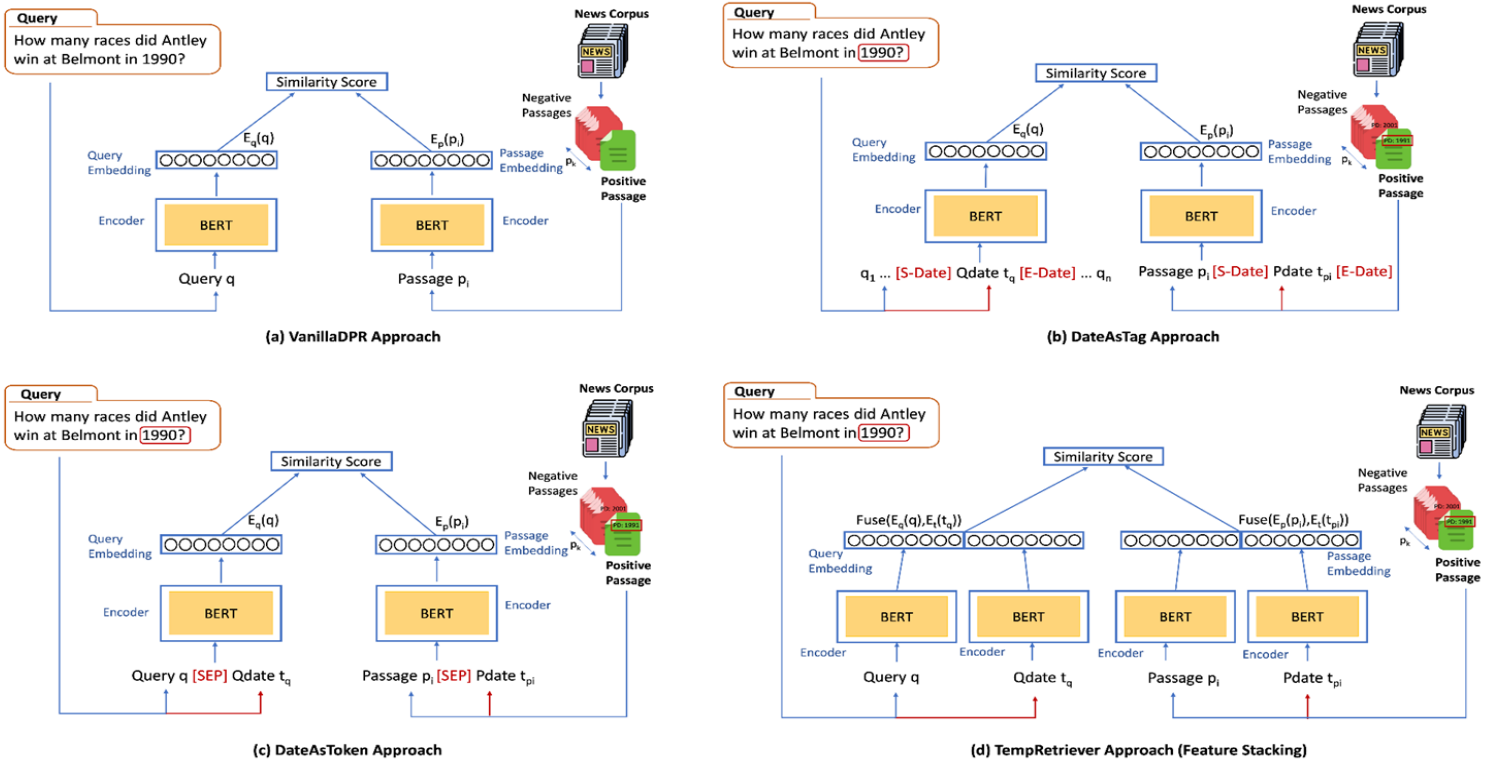
Predict the doc's publication timestamp → task-level temporal supervision.

+155%
over BERT on event-time estimation

Jointly trained on TAMLM + DD with two extra prediction heads.
Beats BERT, RoBERTa & TempoBERT on 7 time-sensitive datasets (event ordering / time estimation / semantic change / QA).

TempRetriever

Core idea: **Temporal retrievers should be also aware of explicit temporal representation**



TempRetriever

QUERY PIPELINE

Query: *"Who was CEO of Microsoft in 2010?"* + **2010**

1. Text Encoder (BERT)
Encodes query text → semantic embedding

2. Temporal Encoder
Encodes timestamp 2010 → temporal embedding

3. Fusion Layer
Combines both embeddings into query representation

DOCUMENT PIPELINE

Document: *"Steve Ballmer was appointed CEO..."* + **2000**

1. Text Encoder
Encodes document text → semantic embedding

2. Temporal Encoder
Encodes timestamp 2000 → temporal embedding

3. Fusion Layer
Combines both embeddings into document representation

SCORING

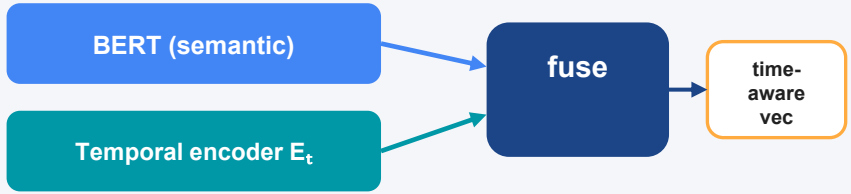
Fusion strategy (e.g., concatenation + learned weights) produces final relevance score by computing similarity between the fused query and document vectors.

TempRetriever Fusion

[Abdallah, Piryani, Wallat, Anand & Jatowt, WSDM 2026]· where TLMs give way to temporally-aware retrieval

Lightweight idea: fuse a learned temporal embedding into a standard dense retriever (DPR) — no specialized pretraining, no architecture surgery.

ARCHITECTURE



FOUR FUSION STRATEGIES

FS	**Feature Stacking** concatenate	VS	**Vector Summation** element-wise add
RE	**Relative Embeddings** query–doc time gap	EWI	**Element-Wise Inter.** element-wise multiply

Goal & Key Insight

Optimization Goal

Ensure relevant, temporally-aligned documents score the highest.

Why Temporal Negatives?

Crucial forces that compel the model to learn precise temporal discrimination.

Core Insight

"Time-based negative sampling teaches the model to distinguish 'relevant but wrong time' from 'right time but irrelevant'."

Training the TempRetriever

TRAINING DATA & SETUP

Training Datasets

- **ArchivalQA:** 532k pairs from NYT (1987–2007)
- **ChroniclingAmericaQA:** 485k pairs from historical papers (1800–1920)

Training Objective

Contrastive Loss using in-batch negatives with a batch size of 32.

CONTRASTIVE EXAMPLES

Positive Example

Query (with timestamp) + relevant document (with matching timestamp)

Negative Examples (Three Types)

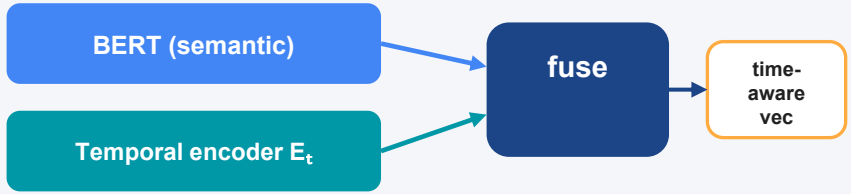
- **Semantic negative:** Topically irrelevant document.
- **Temporal negative:** Relevant document from the *wrong* time period.
- **Mixed negative:** Irrelevant and temporally misaligned.

TempRetriever Fusion

[Abdallah, Piryani, Wallat, Anand & Jatowt, WSDM 2026]· where TLMs give way to temporally-aware retrieval

Lightweight idea: fuse a learned temporal embedding into a standard dense retriever (DPR) — no specialized pretraining, no architecture surgery.

ARCHITECTURE



FOUR FUSION STRATEGIES

FS — **Feature Stacking**
concatenate

VS — **Vector Summation**
element-wise add

RE — **Relative Embeddings**
query–doc time gap

EWI — **Element-Wise Inter.**
element-wise multiply

+6.86%
ArchivalQA R@1

+4.40%
ChroniclingAmericaQA R@1

+9.62%
vs. BiTimeBERT

+5.16%
vs. TS-Retriever

Modular: same fusion boosts BiTimeBERT (+5.12%) & TS-Retriever (+6.17%); zero-shot on NobelPrize; gains carry through to RAG.

Architectural Approaches

Beyond feeding a timestamp — **change HOW the model attends, masks & represents time** so temporal structure is internalized.

Temporal Attention

Rosin & Radinsky, 2022

Time-aware self-attention: scores conditioned on the document's time → time-specific representations.

time-aware attention

Temporal Span Masking

Cole et al., 2023

Salient-span masking targeted at temporal expressions → reasoning over dates & durations.

specialized masking

TALM

Ren et al., 2023

Aligns representations with time-related cues and event structure during training.

time-aligned repr.

SG-TLM

Su et al., 2023

Structure-guided objectives that attend to durations and event structure.

event/duration cues