LocalPIECE: A Locally Executable Framework for News Timeline Construction

Chien-Yu Chu and **Yi-Shin Chen***
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan

Abstract

News timeline construction requires organizing evidence from relevant and peripheral articles, inconsistent participant references, and differences between event and publication dates. We present LocalPIECE, a locally executable extension of PIECE's actor-centered temporal graphs. It combines optional few-shot article weighting, conservative mention and date normalization, redundancy-aware actor selection, and revised graph scoring.

Our evaluation separates backbone development using retained external preprocessing artifacts from integration of the local replacement pipeline. Controlled comparisons show improved date ranking from the graph revisions; fixed diffusion removes encoder training and runs 47 times faster during encoding without changing downstream date scores. Local preprocessing reveals interactions among components rather than uniformly additive improvements.

Experiments on a second topic suggest that sustained structural changes and event onsets may benefit from different temporal scoring strategies. In our evaluation, the complete pipeline processes a news collection in approximately 30 minutes on a consumer laptop, with peak RAM and GPU memory below 3 GB. The results establish local feasibility and clarify how structural change and temporal placement affect date ranking. Selecting a finite timeline without reference-based threshold tuning remains an open challenge.

Keywords: News timeline construction, temporal graphs, few-shot learning, resource-efficient NLP

1 Introduction

A news timeline helps readers follow developments across an article collection. Before generating a summary, a system must decide which dates matter and which source evidence belongs to them. This task is challenging in complex news stories: peripheral reports repeat a topic name, the same participant appears under several names or pronouns, and reports describe earlier events alongside current developments. Publication dates alone do not identify when these events occurred.

PIECE (Hsu et al., 2024) addresses date selection through evolving concept graphs. It detects changes in both global and protagonist-centered graphs. Its concept graphs, actor-local views, and dated evidence provide a framework for analyzing identity, temporal placement, and structural change. Its use of external semantic preprocessing, however, motivates a locally executable alternative for Traditional Chinese news. Replacing that preprocessing also changes the evidence supplied to the graphs; therefore local execution and date-selection quality must be evaluated separately.

LocalPIECE retains the graph-based backbone while revising both its input and its computation. An optional few-shot filter assigns graded article weights. Partial alias/pronoun normalization consolidates supported participant references, and event-time assignment places evidence on explicit or contextually supported dates. A revised actor detector reduces redundant graph anchors. Fixed graph diffusion replaces encoder training; personalized PageRank defines actor-local graphs; and observed-date histories and alternative aggregation rules improve temporal graph scoring.

The study has two stages. We first evaluate backbone revisions with the retained API-derived preprocessing held in place. We then fix this backbone and integrate local evidence processing. Because PIECE's published fixed-threshold policy cannot be uniquely reconstructed, historical comparisons are distinguished from controlled LocalPIECE comparisons.

We investigate whether this graph-based frame-

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 383–392
November 6-7, 2026 ©2026 Association for Computational Linguistics

work can operate locally and how its computational and evidence-processing revisions affect date ranking. Our contributions are a local, source-linked date-ranking pipeline; controlled analysis of actor selection, query encoding, and temporal graph scoring; and an integration study of article influence, mention identity, event time, and output support. The experiments establish feasibility and specific design effects, not superiority over the broader set of timeline systems.

2 Related Work

Timeline summarization combines date selection and evidence organization. Earlier work uses joint date ranking (Tran et al., 2015) and temporally adapted submodular objectives (Martschat and Markert, 2018). Other approaches separate concurrent timelines (Yu et al., 2021), construct event graphs (Li et al., 2021), or use LLM-based clustering (Hu et al., 2024). PIECE tracks dependency-derived concept graphs around recurring protagonists. LocalPIECE extends this framework by revising its preprocessing and graph-based date-ranking components.

Automatically masked expressions have supplied proxy cases for pronoun resolution (Kocijan et al., 2019; Liu et al., 2017; Shen et al., 2021). We use masking to fit small candidate rankers and estimate when bounded rules may act. This connects imperfect heuristic evidence (Bach et al., 2017) with selective prediction and abstention (Geifman and El-Yaniv, 2019), without treating recovery cases as gold coreference or timestamp annotations. The article filter uses SetFit's few-shot classification procedure (Tunstall et al., 2022).

Traditional Chinese has tokenization and parsing resources (Lee and Wang, 2018). The narrower problem here is obtaining compatible, locally executable high-level preprocessing without dedicated mention, temporal-expression, or event-span annotations for the target corpus. The complete weighted pipeline still uses human article labels; it is not claimed to be wholly annotation-free.

3 LocalPIECE

3.1 Task and Graph Backbone

Given query q and retrieved articles $\mathcal{A} = \{a_i\}$, the system returns distinct dates ranked by score, each linked to a nonempty set of complete source sentences. Retrieval is supplied upstream. A sentence's effective date set $\widehat{D}_s$ is estimated by the system and need not equal its true event dates. The candidate domain is the union of retained effective assignments. The task is date ranking with source evidence, not event extraction or event-exclusive evidence selection.

Tokenization, part-of-speech tagging, and dependency parsing identify *conceptual terms*: normalized nouns, proper nouns, and verbs. A direct head–dependent relation contributes an undirected graph edge when both endpoints are retained conceptual terms.

Two graph representations have different roles. A *static query graph*, built from the whole retained collection, represents concept association with the query. A *general temporal graph* accumulates dated dependency evidence. Each selected nominal anchor, called an *actor*, defines a local subgraph of this temporal graph, called an *actor facet*. The general and actor facets are scored for structural change; the static graph supplies query relevance.

Processing assigns article weights, normalizes supported mentions and dates, selects actors, builds temporal facets, and ranks dates from their changes.

3.2 Article Weighting and Normalization

Article influence. The Soft Filter (SF) uses SetFit with `BGE-small-zh-v1.5`. Its four roles are `main`, `secondary`, `brief_mention`, and `none`: primarily about the queried case, substantively related, only mentioning it in the background, or making no meaningful mention, respectively. A binary relatedness classifier supplies the probability of each compatible branch: main/secondary versus brief/none. A flat four-class classifier supplies the relative probabilities within each branch. If $B(y)$ is the branch of role y, their composed probability is

$$\widetilde{P}(y \mid a) = \widehat{P}_R(B(y) \mid a) \frac{\widehat{P}_F(y \mid a)}{\sum_{y' \in B(y)} \widehat{P}_F(y' \mid a)}. \tag{1}$$

Here R and F denote the binary and flat classifiers. With contributions $g(y) = (1, .6, .2, .05)$ in role order, article a_i receives

$$\omega_i = \sum_y g(y) \widetilde{P}(y \mid a_i). \tag{2}$$

These weights attenuate graph evidence rather than delete all peripheral articles. The 16-shot setting uses per-class anchors for both classifiers, except

that none is capped at four. Omitting SF sets all $\omega_i = 1$. Only SF requires dedicated manual article labels.

Mention identity. Alias/pronoun normalization (AR) maps supported mention occurrences to preferred surface forms already present in the corpus. Nominal rules use full names, title–name forms, shorter forms, repeated explicit usage, and article-local uniqueness. Ambiguous matches remain unresolved. For ordinary third-person pronouns, candidates come from local predicate chains, nearby discourse units, and reporting subjects in the preceding one or two sentences. A logistic ranker uses 16 non-lexical features covering syntactic and discourse context. Its compatibility $u_\phi(c \mid m) \in [0, 1]$ is combined with the candidate source's heuristic prior $c_h(m, c) \in [0, 1]$:

$$s(m, c) = .52u_\phi(c \mid m) + .48c_h(m, c). \quad (3)$$

If no ranker can be fitted, the rule score is used alone. The model cannot propose a candidate absent from the rules. Compatibility checks, score thresholds, competition margins, and insufficient support can cause abstention.

Masked explicit subjects provide pseudo-labeled recovery cases for article-grouped fitting and threshold selection. Conservative acceptance requires sufficient pseudo-precision, candidate compatibility, and score margins; otherwise, mentions remain unresolved. Masked-case performance does not guarantee accuracy on natural pronouns.

Event time. The timestamp module (T) recognizes complete and partial dates, relative expressions, and ranges. Bounded contextual candidates come from nearby dated anchors, predicate/content similarity, participant continuity, quotation structure, and opening-event context. Small logistic rankers resolve competing candidates using support, distance, anchor agreement/conflict, publication-time distance, and syntactic features. Explicit dates are masked to form fitting and article-grouped held-out recovery cases. Directly supported rules do not require learned competition.

A non-publication candidate must beat the publication-date hypothesis and pass conflict, cue, and temporal-frame checks. Context can also confirm publication time as the event date. If no date is accepted, publication time remains the fallback. For multi-date sentences, clause boundaries, neighboring date anchors, and compatible predicate/argument structure produce approximate event scopes. Graph edges are projected to the corresponding date scopes, rather than copying every sentence edge to every date. These scopes distribute graph evidence; they are not extracted event records. Final evidence retains whole sentences rather than selecting their highest-weighted event spans.

Before graph construction, a corpus-derived year frame excludes remote historical records: each article votes at most once per resolved year; years supported by at least 5% of articles define a span expanded by one guard year on either side. This admission rule uses no reference dates. AR/T implement partial normalization with rules, masked-case fitting, and abstention rather than dedicated linguistic labels. We study how this implementation changes graph evidence; individual internal heuristics are not independently validated.

3.3 Actor Selection and Local Facets

PIECE ranks noun frequencies separately in subject and object subtrees, combines these ranks, and selects five anchors. Long phrases can credit many overlapping words, frequent generic words can dominate, and several anchors can have redundant contexts. LocalPIECE addresses these issues in three steps.

Bounded mention credit. A subject/object occurrence m contributes total credit $\omega(m)b_r$, where $b_{\text{subj}} = 1$ and $b_{\text{obj}} = .7$. Candidates are nouns/proper nouns of at least two characters. The nominal head and attached compound/name or appositional nouns share this credit; punctuation blocks expansion, loose nominal modifiers are excluded, and coordinated heads receive separate occurrence credits. For member set $M(m)$,

$$\text{cred}(m, w) = \omega(m)b_r \frac{\pi_m(w)\,\text{idf}(w)^{.5}}{\sum_{u \in M(m)} \pi_m(u)\,\text{idf}(u)^{.5}}. \quad (4)$$

The head prior is 1 and attached-member prior .6. For example, suppose a subject occurrence has weight 1 and two eligible nouns: head "bureau" and compound member "health." With equal IDF factors, their credits are $1/1.6 = .625$ and $.6/1.6 = .375$. Adding another attached noun redistributes this same total rather than creating extra subject evidence. An object occurrence instead shares total credit .7. With total weighted mention count N and mention count n_w containing w,

the IDF-like factor is $1 + \log[(1 + N)/(1 + n_w)]$. Summed subject and object credits become percentile ranks R_s, R_o over positive values; absent evidence has rank zero. Their role score is $C(w) = (R_s(w) + .7R_o(w))/1.7$.

Genericity and redundancy. Let $p_w(u)$ be the normalized article-weighted counts of nominal partners u sharing eligible mentions with w. Define $H(w)$ as partner entropy divided by the log of the partner count, $B(w)$ as $\log(1 + |\,\text{Part}(w)|)$ divided by the largest such value, and $r_a(w)$ as the fraction of mentions in which w appears without another eligible noun. Undefined ratios are zero. The candidate score is

$$\begin{aligned} G(w) &= H(w)B(w)[1 - r_a(w)], \\ s_{\text{act}}(w) &= C(w)e^{-.35G(w)}. \end{aligned} \tag{5}$$

This discounts a word used diffusely with many different nominal partners. It does not classify the word's semantic importance.

Pair similarity $\sigma(w, u)$ is the larger of literal surface similarity and a weighted mixture of surface overlap (.45), mention co-membership (.25), and context cosine (.30). Co-membership divides the shared-mention count by the geometric mean of mention counts; context vectors count grammatical roles and governing predicates. Surface matching uses exact identity, conservative containment, or high character-bigram overlap. Normalize s_{act} by its maximum to obtain $\bar{s}$. Starting with $S = \varnothing$, select five actors sequentially by maximal marginal relevance:

$$\arg\max_{w \notin S} \left[.85\bar{s}(w) - .15 \max_{u \in S} \sigma_+(w, u) \right], \tag{6}$$

where $\sigma_+ = \max\{0, (\sigma - .75)/.25\}$ and the empty-set penalty is zero. The first actor therefore maximizes centrality; later actors can favor a slightly lower-scoring noun if it overlaps less with those already selected. "Divergent" thus means less redundant lexical/syntactic usage, not opposing views.

Temporal facets. Let $O_t(u, v)$ be dependency evidence assigned to date t, with article weights applied. The general graph has edge weights

$$G_{0,t}(u, v) = \sum_{t' \le t} e^{-.005(t-t')} O_{t'}(u, v), \tag{7}$$

where elapsed time is measured in calendar days. For actor w, personalized PageRank with damping .85 localizes the current graph. Nodes are retained in decreasing PageRank mass, targeting .80 cumulative mass, with at least five nodes when available and a hard cap of 30. The cap can prevent reaching the mass target. The induced subgraph is $G_{w,t}$; an absent actor uses a singleton fallback. These overlapping facets are views of the same graph, not independent evidence sources.

3.4 Query Representation and Temporal Scoring

Static relevance representation. The query graph contains conceptual terms and a query node. Direct dependency co-occurrence supplies Dice connectivity,

$$A^q_{uv} = \frac{2\,\text{freq}(\langle u, v \rangle)}{\text{freq}(u) + \text{freq}(v)}. \tag{8}$$

Query–keyword edges have connectivity 1. Edge distance is $.1 - \log_2 A^q_{uv}$; nodes within the mean reachable shortest-path distance from the query form its ego graph. Initial features X use signed hashing of article incidence into 512 dimensions: term rows represent articles containing them, and the query row represents articles containing its keywords. With self-loops and degree matrix $\widetilde{D}$, fixed diffusion gives

$$\begin{aligned} S &= \widetilde{D}^{-1/2}(A + I)\widetilde{D}^{-1/2}, \\ \Phi &= .75X + .25SX. \end{aligned} \tag{9}$$

The row ϕ_v represents a node, and this single diffusion step replaces the trainable graph autoencoder (GAE). The static graph uses the whole collection, so the method is retrospective, not an online forecasting system.

Structural signature. For each facet $r \in \{0\} \cup \mathcal{P}$, form the unnormalized Laplacian $L_{r,t} = D_{r,t} - A_{r,t}$. Its largest eigenvalues give signature $\mathbf{z}_{r,t}$, with four coordinates for the general facet and two for an actor facet. Eigenvalues are invariant to node ordering, allowing comparisons across changing facets. However, cospectral graphs and changes in evidence volume limit their interpretation as structural changes.

Following Laplacian Anomaly Detection (LAD) (Huang et al., 2020), L2-normalized historical signatures form columns of a history matrix $C^{(h)}_{r,t}$ for the previous h observed evidence dates, $h \in \{7, 14\}$. Its first left singular vector $u^{(h)}_1$ represents the recent typical direction. The

SVD is on this short history matrix, not the graph Laplacian. Window scores are

$$z_r^{(h)}(t) = 1 - \left|\cos(\mathbf{z}_{r,t}, u_1^{(h)})\right|. \quad (10)$$

The absolute value removes the arbitrary singular-vector sign. A zero-vector comparison has score zero. The selected raw signal is $z_r(t) = \max_h z_r^{(h)}(t)$. Only dates with observed evidence update the history; decay alone does not introduce a new normalized structural direction.

For onset-sensitive scoring, take positive change *within each window* before fusing them:

$$\Delta z_r(t) = \max_h [\, z_r^{(h)}(t) - z_r^{(h)}(t_h^-) \,]_+. \quad (11)$$

Here t_h^- is the previous scoreable observed date for that window; its first scoreable value uses an implicit zero predecessor. This is not the first difference of the already fused raw score.

Raw LAD captures an unusual structural state, whereas positive ΔLAD emphasizes its onset.

3.5 Query Gating, Aggregation, and Output

Ordinary PageRank on the general temporal graph identifies concepts whose centrality increased from the previous observed state: $\widehat{V}_t = \{w : \mathrm{pr}_t(w) > \mathrm{pr}_{t^-}(w)\}$. Their mean query association is

$$a(t) = \frac{1}{|\widehat{V}_t|} \sum_{w \in \widehat{V}_t} \cos(\phi_q, \phi_w). \quad (12)$$

Missing representations contribute zero; an empty set gives $a(t) = 0$. Only the general score is gated: $g_t = z_0(t)\mathbf{1}[a(t) > 0]$. Actor scores remain eligible. The two retained aggregation views are

$$Z_{\text{sum}}(t) = g_t + \sum_{w \in \mathcal{P}} z_w(t), \quad (13)$$

$$Z_{\max}(t) = \max\{g_t, \max_{w \in \mathcal{P}} z_w(t)\}. \quad (14)$$

The same operations apply to ΔLAD. Sum combines responses across facets, while Max highlights dates with strong changes in individual facets. Max is selected for the primary configuration, while Sum is retained for comparison. Dates with retained evidence are ranked by Z, with complete assigned source sentences attached. Anomaly scores do not directly measure event importance and are not necessarily comparable across graphs or topics.

Post-score support rules restrict eligible dates without changing graphs or scores. SAk requires k distinct supporting articles; SSk requires k distinct supporting sentences.

These counts do not necessarily represent independent news sources. These case-motivated rules are applied after scoring; sensitivity analysis does not establish an automatic threshold-selection procedure.

4 Experimental Setup

4.1 Data and Metrics

We use two Mandarin collections from PIECE (Hsu et al., 2024). En-En (恩恩案) corresponds to *COVID-19 Patient Tragedy* topic; the Lin Chih-chien thesis-plagiarism controversy (林智堅論文抄襲案) is its *Political Plagiarism Allegations* topic, abbreviated *Plagiarism*. We use the latter name below.

The En-En collection contains 731 articles and 8,271 sentences published between 14 January and 21 August 2022. The Plagiarism collection contains 540 articles and 5,052 sentences from 5 July to 21 August 2022.

Inherited and additional references. We use the retained PIECE reference dates, identified as Yahoo Taiwan human labels in our source audit (Hsu et al., 2024). En-En has 15 unique reference dates after deduplication. We exclude 13 May because its pediatric-policy development cannot be temporally linked to the case in the available corpus, leaving 14 dates. Plagiarism has 13 unique reference dates; excluding 24 August, which falls outside the collection, leaves 12. These accessible counts are distinguished from the 18 selected dates reported in PIECE's table.

We constructed two additional En-En references using the Chinese Wikipedia page 恩恩事件 (Wikipedia contributors, 2026). ChatGPT 5.4 generated essential and extensive event lists, which we checked against the corpus, removing unsupported events without adding new ones. The resulting references contain 13 and 31 dates, respectively. They represent different editorial perspectives rather than nested timelines and were completed before method development. Macro metrics average results across the three En-En references to assess sensitivity to editorial perspective.

Date-level evaluation. A predicted date is correct if it appears in the evaluated reference date

set G. We evaluate exact calendar dates, without a tolerance window or matching associated event descriptions. The metrics do not score intended-event recovery or penalize off-topic sentences. Multiple reference entries on one date collapse to one target. Let $\hat{t}_k$ be the date at rank k in decreasing score order (ties use ascending date), and $H_k = |\{\hat{t}_1, \ldots, \hat{t}_k\} \cap G|$. Then

$$P@k = H_k/k, \qquad R@k = H_k/|G|,$$
$$\text{AP} = \frac{1}{|G|} \sum_{k:\hat{t}_k \in G} P@k. \tag{15}$$

AP measures ranking quality, while Recall@k measures coverage at fixed budgets ($k = 5, 10, 20$ for En-En; $5, 10, 15$ for Plagiarism). Missing or filtered reference dates remain in the denominator.

For a score-thresholded date subset B, precision is $|B \cap G|/|B|$, recall is $|B \cap G|/|G|$, and F1 is $2PR/(P+R)$, with zero for an empty or zero-hit subset. Sweeping observed score thresholds gives *Oracle Best F1* (OF_1); ties favor higher precision, fewer dates, then a higher threshold. Because OF_1 uses reference labels to select the threshold, it measures attainable performance rather than a deployable selection policy.

4.2 Comparison and Development Protocol

PIECE reports precision, recall, and F1 using a fixed threshold of .0005, but its complete selection policy cannot be reconstructed (Hsu et al., 2024). We therefore distinguish controlled backbone comparisons, local preprocessing experiments, and historical PIECE output comparisons.

Backbone development retains PIECE-derived GPT-3 statements and Articut date artifacts. Actor/LAD comparisons use the original GAE policy, followed by corrected encoder comparisons. Graph and aggregation experiments then use the selected representation. Finally, we fix the backbone and replace external preprocessing with the local pipeline.

We re-evaluate retained PIECE scores on the corrected references within their original date domains. Because these domains differ from Local-PIECE's, archive metrics are historical diagnostics rather than matched comparisons. Zero-padding is avoided because it can artificially alter AP.

En-En development explored 29 decision points and 104 options without a separate validation topic. For Plagiarism, we first apply the En-En-selected backbone without target SF labels, then examine raw versus positive ΔLAD and a case-motivated support rule. These analyses are exploratory rather than independently validated configuration choices.

CKIP provides tokenization and POS tagging (Ma and Chen, 2003), while spaCy `zh_core_web_trf` provides dependency parsing. Experiments run on an i7-11800H CPU, RTX 3050 GPU (4 GB VRAM), and 8 GB RAM.

5 Results and Analysis

5.1 Backbone Development with Retained External Support

Actors and anomaly interpretation. Table 1(a) separates detector choice from raw/ΔLAD. Under either scorer, the divergent detector increases AP; under either detector, raw LAD ranks better on this sustained storyline. With raw LAD, changing actors raises AP from .245 to .368, while oracle F1 changes only from .417 to .426. The improvement is mainly in date ranking, with limited gains in oracle F1.

Paired article-level bootstrap diagnostics support reduced redundancy: BGE-base-zh-v1.5 contextual centroid-cosine dissimilarity increases by .0158 (95% CI [.0038,.0238]), and publication-date Jensen–Shannon divergence by .1139 ([.0841,.1716]). These do not establish viewpoint diversity.

Replacing encoder training. The proxy evaluation holds out 5% of eligible query edges for validation and 10% for test, preserving connectivity, query-incident, and degree-one supporting edges. Equal negative sets exclude *all* observed positives, including held-out edges. Inner products rank edges; validation AP selects GAE checkpoints.

Fixed diffusion is 47 times faster than GAE during encoding (Table 2), while producing identical downstream date AP and oracle F1. Although diffusion achieves higher edge-prediction AP, both encoders produce identical date-ranking results. Its main advantage is reduced computational cost.

Temporal graphs and aggregation. Table 1(b) uses a common Base with decay .05, actor ego graphs bounded by mean reachable distance from the anchor, and calendar-day LAD histories. D changes decay to .005; PPR replaces the ego rule; O restricts history to observed evidence dates.

Configuration	Date AP	R@5	R@20	OF_1	Macro AP	Macro OF_1
(a) Actor/scorer development: common pre-optimization GAE policy						
Original detector, ΔLAD	.181	.071	.286	.267	.226	.321
Original detector, raw LAD	.245	.071	.214	.417	.289	.458
Divergent detector, ΔLAD	.303	.143	.357	.313	.332	.363
Divergent detector, raw LAD	.368	.143	.286	.426	.391	.459
(b) Graph development: fixed diffusion, divergent actors, raw LAD						
Base, Sum	.362	.143	.357	.408	.384	.453
+D, Sum	.441	.143	.500	.489	.460	.511
+PPR, Sum	.449	.214	.500	.486	.451	.501
+O, Sum	.404	.214	.357	.465	.432	.493
+D+O+PPR, Sum	.578	.214	.714	.667	.530	.588
+D+O+PPR, Max	.632	.286	.643	.720	.548	.614
Historical reference: its own stored date domain and unreconstructed selection policy						
PIECE archive (re-evaluated)	.527	.286	.714	.606	.379	.477

Table 1: En-En backbone development with retained external preprocessing. Panels (a) and (b) have different encoder policies; they are separate controlled comparisons. D: decay .005; PPR: actor-local PageRank facets; O: observed-date history. OF_1: Oracle Best F1. The historical row is not a new PIECE execution. These are not full local-preprocessing results.

Encoder	Test edge AP	Encoding time (s)	Date AP	OF_1
Corrected GAE	.925	77.04	.362	.408
Fixed diffusion	.951	1.64	.362	.408

Table 2: Controlled query-encoder comparison. Edge AP is an auxiliary proxy; date metrics evaluate the downstream task. Runtime covers encoding, not the complete pipeline. GAE executes 300 epochs; its selected checkpoint is the first evaluation epoch.

Each change improves its AP and oracle F1. Together they reach .578 AP and .667 oracle F1 with Sum. The combined diagnostics also exceed the re-evaluated archive's AP and oracle F1, while the archive retains higher R@5. That historical comparison remains conditional on the different output domains and unknown original selection policy.

Max on the same backbone raises AP to .632 and oracle F1 to .720, with R@5 matching the archive. R@20 falls from .714 under Sum to .643 under Max. Max's oracle subset contains 11 dates, versus 13 for Sum and 19 for the archive. These results motivate a selective Max view while preserving Sum for broader accumulation. These are development diagnostics, not universal configuration preferences.

5.2 Replacing External Preprocessing with the Local Pipeline

The selected backbone is now fixed. Table 3 changes article influence, mention identity, and temporal placement. Sum gives a readable assembly path; Max is then evaluated on the same final evidence. Unlike the previous experiments, these conditions use local sentence-based processing.

Weighting changes influence, not necessarily AP. The 16-shot flat role classifier has class-support-weighted F1 .827±.003 over 20 seeds (mean ± standard error). This auxiliary classification result does not establish probability calibration or downstream ranking quality.

SF alone reduces AP under both publication and local event-time input. Its effect is nevertheless measurable before date scoring: brief/none articles contribute 22.8% of unweighted statement mass, but 8.2% after weighting. These modeled influence shares do not measure gold noise. Down-weighting changes edges, actors, and graph histories; isolated ranking gains need not follow.

Identity and temporal placement interact. Adding AR after SF raises AP from .285 to .344 with publication dates and from .341 to .498 with event-time input. Without SF, event-time+AR reaches only .441. These results suggest interactions among article weighting, identity normalization, and temporal placement, rather than independent additive improvements. They do not establish the intrinsic accuracy of the normalization modules.

Event-time input makes all 14 main-reference

Temporal input	Processing	Aggregation	Date AP	OF_1	Macro AP	Macro OF_1
Publication dates	none	Sum	.376	.571	.348	.495
Publication dates	SF	Sum	.285	.438	.291	.452
Publication dates	SF+AR	Sum	.344	.516	.349	.504
Local event-time	none	Sum	.431	.563	.444	.543
Local event-time	SF	Sum	.341	.444	.358	.450
Local event-time	AR	Sum	.441	.529	.457	.534
Local event-time	SF+AR	Sum	.498	.588	.489	.557
Local event-time	SF+AR	Max	.499	.667	.455	.590
Historical reference: stored domain, unreconstructed original selection policy						
PIECE archive	—	archive	.527	.606	.379	.477

Table 3: Local-pipeline integration under the fixed graph/LAD backbone. SF: soft article weighting; AR: alias/pronoun normalization. Changes to temporal input can change the candidate domain. The assembly path is descriptive, not evidence of additive component gains. Archive metrics retain its own domain.

dates available, compared with 10 under publication-only placement. It represents 8,586 event projections over 8,271 sentences, including 198 sentences with more than one effective date. Publication-date concentration falls to 79.0%; this includes both supported publication-date assignments and fallback. These figures describe representation, not timestamp-error rates.

Local output quality and actor evidence. Sum and Max yield nearly identical main AP (.498/.499), but Max increases oracle F1 from .588 to .667 and macro oracle F1 from .557 to .590; macro AP falls from .489 to .455. Thus Max's benefit in the integrated pipeline concerns the attainable finite subset rather than a uniform ranking improvement. Under Sum, retaining only actor facets gives AP .506 and 10 of 14 main dates in the top 20; general-only gives .333 and seven hits. These overlapping facets are consequential but are not independent sources.

The full local Max configuration has lower AP than the historical archive (.499 versus .527), but higher oracle F1 (.667 versus .606). These values are not directly comparable because the candidate domains differ.

This is a historical multi-metric trade-off, not a matched-policy verdict on the two models. The .632 AP in Table 1 uses retained external preprocessing. Replacing this support costs some ranking quality while retaining useful finite-set separation.

5.3 Second Topic and Output Support

The raw/ΔLAD contrast has a prior hypothesis: raw LAD should favor sustained departures, whereas positive change should emphasize episode onset. ΔLAD belongs to the original backbone; selecting raw LAD on En-En was not assumed to generalize across reporting patterns. When the transferred raw core underperformed on Plagiarism, we examined this existing alternative and reported both scorers (Table 4). With the same 61-date domain, ΔLAD raises AP from .265 to .380 and oracle F1 from .408 to .558. This supports the state-versus-onset hypothesis, although further validation is needed to determine which scorer is preferable for unseen topics.

The case-motivated SS10 rule retains 37 dates, with AP .524 and oracle F1 .649, preserving all 12 reference dates. The threshold was selected before sensitivity analysis and was not replaced by its best-performing alternative. SS10–SS18 give the same .524/.649 diagnostics; stricter thresholds can improve them further. Retaining SS10 illustrates support separation rather than an optimum at ten. The archive remains higher at R@15 (.583 versus .500).

Support is less cleanly separated on En-En. Table 5 shows that two to five supporting articles improve Max ranking while retaining all main-reference dates; ten retains only 12 of 14. A late, substantive 16 August one-dollar compensation filing has only two articles/two sentences and is removed by SA10. A sparse event can matter without adding enough new structure to change a mature graph strongly. Conversely, small new facets can produce large structural deviations from little evidence. Support is therefore an uncertainty cue, not a universal substitute for importance.

5.4 Computational Resources

Full En-En processing takes approximately 30 minutes, including SF label production timed separately in its notebook and added to the pipeline run-

Condition (AR+event-time+Max; no target SF)	Dates	Date AP	OF_1	R@5	R@15
Historical reference					
PIECE archive (re-evaluated)	27	.472	.606	.250	.583
Frozen core					
Raw LAD	61	.265	.408	.167	.167
Hypothesis-motivated scorer comparison					
Positive ΔLAD	61	.380	.558	.167	.417
Preselected support rule					
Positive ΔLAD+SS10	37	.524	.649	.333	.500

Table 4: Political Plagiarism Allegations. Dates is the eligible domain. The scorer comparison investigates an existing hypothesis; SS10 is a preselected post-score support restriction. Neither establishes automatic topic adaptation.

Max condition	Coverage	Date AP	OF_1
No support filter	14/14	.499	.667
SA2	14/14	.611	.690
SA5	14/14	.618	.690
SA10	12/14	.572	.667

Table 5: En-En article-support sensitivity. Coverage is reference-date retention before ranking, not top-ranked recall. Removed reference dates remain in metric denominators. The sweep characterizes support/coverage trade-offs; it is not used to replace the preselected threshold.

time. POS/dependency parsing dominates. With expensive outputs precomputed, remaining computation takes six minutes. Peak RAM is 2.2 GB and GPU memory 2.5 GB. These establish local feasibility, not a matched total-cost advantage over PIECE.

6 Limitations and Conclusion

LocalPIECE ranks structural changes in reported evidence, not calibrated event importance. Sparse consequential events and cold-start expansion can mislead ranking. AR/T are exploratory heuristic integrations: downstream changes characterize this implementation, not the individual efficacy of every rule or intrinsic mention/date accuracy. Internal weights and event scopes distinguish evidence contributions, but this distinction is not carried into final sentence selection. Relevant developments may be included alongside substantial off-topic material; we do not require event-exclusive evidence. Date hits do not certify recovery of the intended event. Development-topic selection, two-topic coverage, and the historical PIECE comparison constrain the conclusions.

Controlled comparisons nevertheless support actor and graph revisions, and fixed diffusion reduces encoder cost without downstream loss in the tested comparison. Local execution exposes interactions among article influence, identity, and time. The principal remaining step is a reference-free output policy. Hypothesis-consistent scorer contrasts and support sensitivity do not yet prescribe how to select a scorer, evidence unit, or finite date set on an unseen topic. The selection policy requires held-out validation.

References

Stephen H. Bach, Bryan He, Alexander Ratner, and Christopher Ré. 2017. Learning the structure of generative models without labeled data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 273–282. PMLR.

Yonatan Geifman and Ran El-Yaniv. 2019. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159. PMLR.

Tz-Huan Hsu, Li-Hsuan Chin, Yen-Hao Huang, and Yi-Shin Chen. 2024. PIECE: Protagonist identification and event chronology extraction for enhanced timeline summarization. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pages 870–879. Association for Computing Machinery.

Qisheng Hu, Geonsik Moon, and Hwee Tou Ng. 2024. From moments to milestones: Incremental timeline summarization leveraging large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7232–7246, Bangkok, Thailand. Association for Computational Linguistics.

Shenyang Huang, Yasmeen Hitti, Guillaume Rabusseau, and Reihaneh Rabbany. 2020. Laplacian change point detection for dynamic graphs. In

Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '20, pages 349–358. Association for Computing Machinery.

Vid Kocijan, Oana-Maria Camburu, Ana-Maria Cretu, Yordan Yordanov, Phil Blunsom, and Thomas Lukasiewicz. 2019. WikiCREM: A large unsupervised corpus for coreference resolution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4303–4312, Hong Kong, China. Association for Computational Linguistics.

Yen-Hsuan Lee and Yih-Ru Wang. 2018. 繁體中文依存句法剖析器 (Traditional Chinese Dependency Parser) [in chinese]. In *Proceedings of the 30th Conference on Computational Linguistics and Speech Processing (ROCLING 2018)*, pages 61–75, Hsinchu, Taiwan. The Association for Computational Linguistics and Chinese Language Processing.

Manling Li, Tengfei Ma, Mo Yu, Lingfei Wu, Tian Gao, Heng Ji, and Kathleen McKeown. 2021. Timeline summarization based on event graph compression via time-aware optimal transport. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6443–6456, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ting Liu, Yiming Cui, Qingyu Yin, Wei-Nan Zhang, Shijin Wang, and Guoping Hu. 2017. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 102–111, Vancouver, Canada. Association for Computational Linguistics.

Wei-Yun Ma and Keh-Jiann Chen. 2003. Introduction to CKIP chinese word segmentation system for the first international chinese word segmentation bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 168–171, Sapporo, Japan. Association for Computational Linguistics.

Sebastian Martschat and Katja Markert. 2018. A temporally sensitive submodularity framework for timeline summarization. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 230–240, Brussels, Belgium. Association for Computational Linguistics.

Ming Shen, Pratyay Banerjee, and Chitta Baral. 2021. Unsupervised pronoun resolution via masked noun-phrase prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 932–941, Online. Association for Computational Linguistics.

Giang Tran, Eelco Herder, and Katja Markert. 2015. Joint graphical models for date selection in timeline summarization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1598–1607, Beijing, China. Association for Computational Linguistics.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. In *Proceedings of the Second Workshop on Efficient Natural Language and Speech Processing (ENLSP-II) at NeurIPS 2022*.

Wikipedia contributors. 2026. 恩恩事件 [en-en incident]. Wikipedia, the Free Encyclopedia. Accessed 3 August 2026.

Yi Yu, Adam Jatowt, Antoine Doucet, Kazunari Sugiyama, and Masatoshi Yoshikawa. 2021. Multi-TimeLine summarization (MTLS): Improving timeline summarization by generating multiple summaries. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 377–387, Online. Association for Computational Linguistics.