Journal of Biomedical Informatics

Contextualized Medication Event Extraction with Levitated Markers

--Manuscript Draft--

Manuscript Number:	
Article Type:	VSI: Clin NLP secondary use
Keywords:	clinical NLP; text mining; context classification; event extraction; levitated markers
Corresponding Author:	Jake Vasilakes The University of Manchester Manchester, UNITED KINGDOM
First Author:	Jake Vasilakes
Order of Authors:	Jake Vasilakes
	Panagiotis Georgiadis
	Nhung T.H. Nguyen, PhD
	Makoto Miwa, PhD
	Sophia Ananiadou, PhD
Abstract:	Automatic extraction of patient medication histories from free-text clinical notes can increase the amount of relevant information to clinicians for developing treatment plans. In addition to detecting medication events, clinical text mining systems must also be able to predict event context, such as negation, uncertainty, and time of occurrence, in order to construct accurate patient timelines. Towards this goal, we introduce Levitated Context Markers (LCMs), a novel transformer-based model for contextualized event extraction. LCMs are an adaptation of levitated markers --- originally developed for relation extraction--- that allow pretrained transformer models to utilize global input representations while also focusing on event-related subspans using a sparse attention mechanism. In addition to outperforming a strong baseline model on the Contextualized Medication Event Dataset, we show that LCMs' sparse attention can provide interpretable predictions by detecting relevant context cues in an unsupervised manner.
Suggested Reviewers:	
Opposed Reviewers:	

National Centre for Text Mining
Department of Computer Science
The University of Manchester, Manchester, UK

January 6, 2023

Dear Editors,

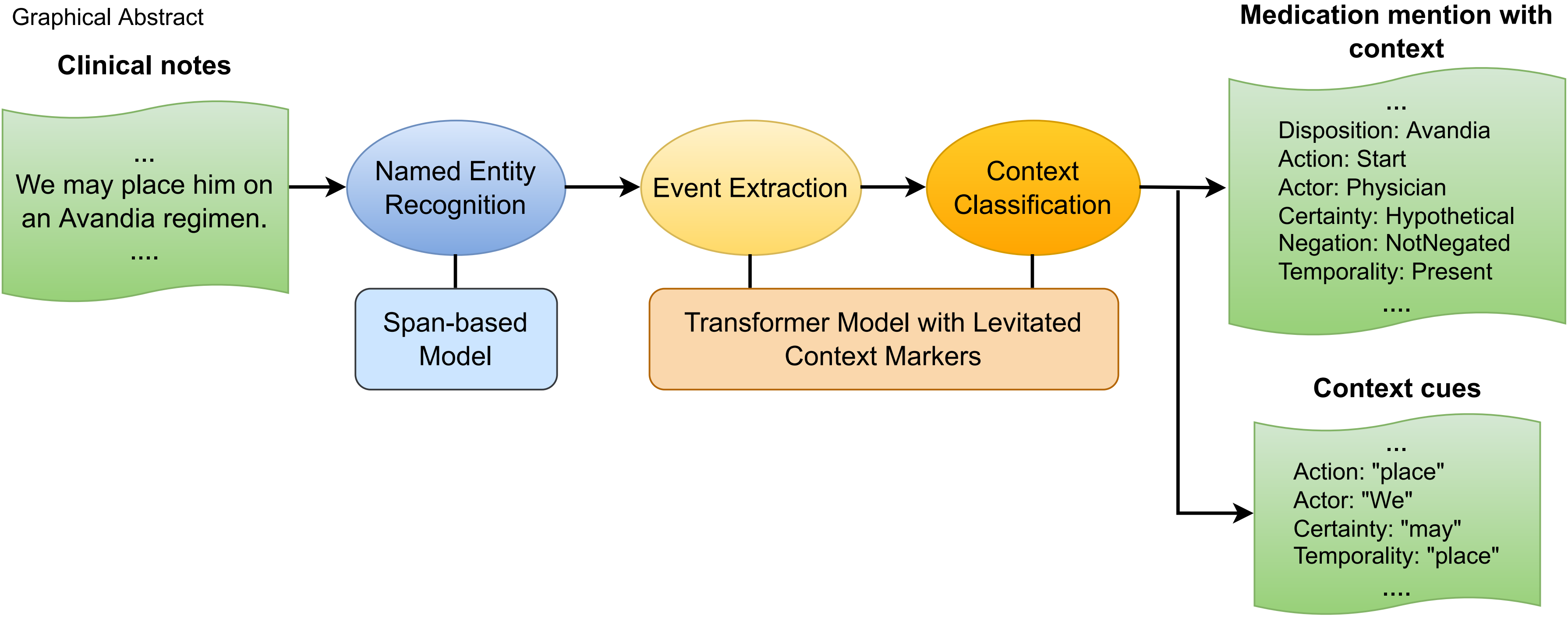
We are pleased to submit our original research article "Contextualized Medication Event Extraction with Levitated Markers" to the Special Issue on Clinical Natural Language Processing for Secondary Use Applications of the Journal of Biomedical Informatics.

This manuscript extends research related to Track 1 of the 2022 n2c2 Shared Task on contextualized medication event extraction. We introduce Levitated Context Markers (LCMs), a new transformer-based method for the extraction of medication change events and classification of their context. LCMs allow the model to utilize global input representations while also focusing on event-related subspans using a sparse attention mechanism. Experiments on the Contextualized Medication Extraction Dataset show that LCMs outperform a strong baseline model for both event extraction and context classification. Additionally, we show that the sparse attention mechanism is able to detect relevant cues in the input —such as negation triggers, modals, and verbs of change— in an unsupervised manner, lending interpretability to the model predictions.

We declare that this manuscript is original, unpublished, and is not under consideration for publication elsewhere. Thank you for taking the time to consider our manuscript for publication. We look forward to hearing from you.

Sincerely,
Jake Vasilakes, Panagiotis Georgiadis, Nhung T.H. Nguyen, Makoto Miwa, and Sophia Ananiadou

Graphical Abstract



Statement of Significance

Contextualized Medication Event Extraction with Levitated Markers

Jake Vasilakes, Panagiotis Georgiadis, Nhung T.H. Nguyen, Makoto Miwa, Sophia Ananiadou

Problem

Effective clinical text mining systems must be able to extract both events and their context.

What is already known

While previous works have studied the classification of event context, the datasets used provide detailed annotations of event structure and cue words. These annotations are time-consuming to obtain, limiting the development of text mining systems.

What this paper adds

We propose Levitated Context Markers (LCMs) for contextualized event extraction. LCMs outperform a strong baseline on the Contextualized Medication Event Dataset, which contains only high-level context annotations and no annotation of cue words. LCMs are also able to detect relevant cue spans without explicit supervision, which adds interpretability to the model's predictions.

Declaration of interests

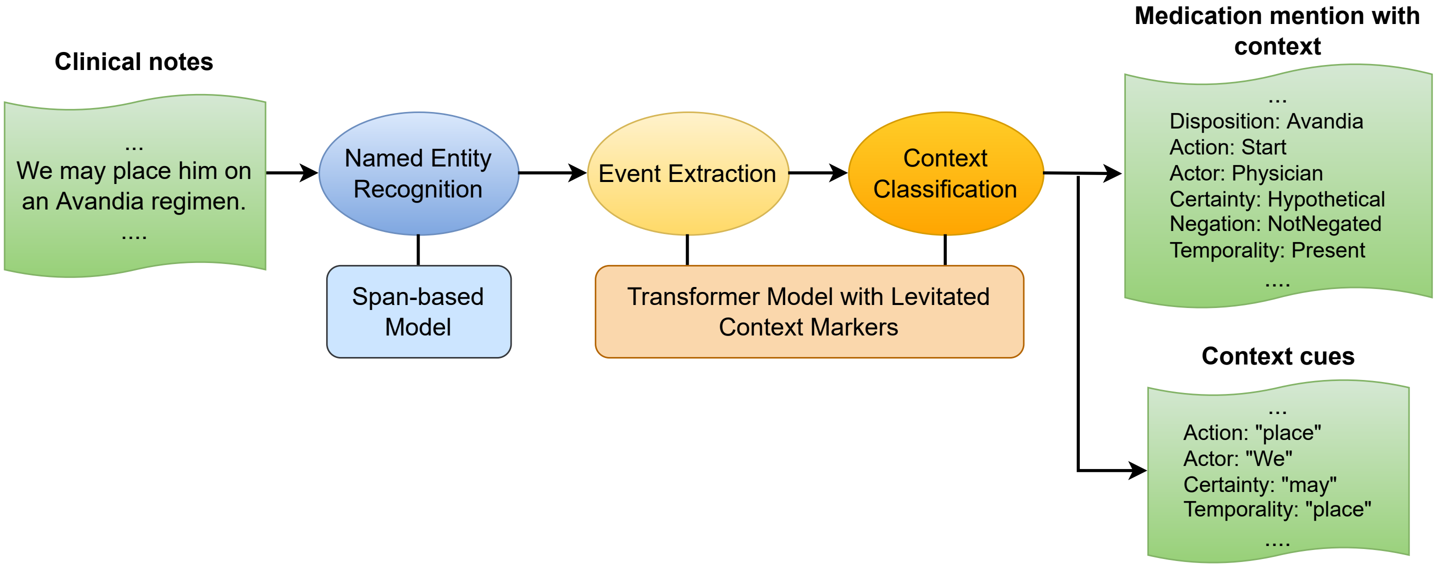
☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Graphical Abstract

Contextualized Medication Event Extraction with Levitated Markers

Jake Vasilakes, Panagiotis Georgiadis, Nhung T.H. Nguyen, Makoto Miwa, Sophia Ananiadou



Highlights

Contextualized Medication Event Extraction with Levitated Markers

Jake Vasilakes, Panagiotis Georgiadis, Nhung T.H. Nguyen, Makoto Miwa, Sophia Ananiadou

- We propose Levitated Context Markers (LCMs) a new method for contextualized event extraction in clinical text.
- LCMs use a sparse attention mechanism which is able to automatically detect relevant cue spans in an unsupervised manner.

Contextualized Medication Event Extraction with Levitated Markers

Jake Vasilakes[a,1], Panagiotis Georgiadis[a,1], Nhung T.H. Nguyen[a], Makoto Miwa[b,c], Sophia Ananiadou[a,c,d,*]

[a] *National Centre for Text Mining, The University of Manchester, Manchester, UK*
[b] *Toyota Technological Institute, Nagoya, Japan*
[c] *Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan*
[d] *Alan Turing Institute, London, UK*

Abstract

Automatic extraction of patient medication histories from free-text clinical notes can increase the amount of relevant information to clinicians for developing treatment plans. In addition to detecting medication events, clinical text mining systems must also be able to predict event context, such as negation, uncertainty, and time of occurrence, in order to construct accurate patient timelines. Towards this goal, we introduce Levitated Context Markers (LCMs), a novel transformer-based model for contextualized event extraction. LCMs are an adaptation of levitated markers —originally developed for relation extraction— that allow pretrained transformer models to utilize global input representations while also focusing on event-related subspans using a sparse attention mechanism. In addition to outperforming a strong baseline model on the Contextualized Medication Event Dataset, we show that LCMs' sparse attention can provide interpretable predictions by detecting relevant context cues in an unsupervised manner.

Keywords: clinical NLP, text mining, context classification, event extraction, levitated markers

1. Introduction

Electronic Health Records (EHRs) provide clinicians access to detailed patient histories at the point of care, improving outcomes by informing treatment options. The structured data contained in EHRs is easy to search, but vital information often lies exclusively in free-text clinical notes, which are time-consuming to review [1]. Clinical text mining aims to automatically extract structured information from clinical notes in order to augment EHRs, increasing the availability of patient data to clinicians.

Clinical text mining is well-studied, and relationship or event extraction is a key task [2]. However, the *context* surrounding the extracted events is equally important for downstream reasoning and interpretation. For example, negation and certainty determine the factuality of an event [3], and event temporality (e.g., past, present, future) is a prerequisite for constructing patient timelines [4]. The new Contextualized Medication Event Dataset (CMED) —released as part of the 2022 n2c2 shared task [5]— brings multiple contexts together into a single dataset, providing annotations of 5 context dimensions (Action, Actor, Certainty, Negation, and Temporality) relating to medication change events in clinical notes.

Using CMED as our test bed, we propose a new method for contextualized event extraction called Levitated Context Markers (LCMs). LCMs are a variant of levitated markers, which were originally developed for NER and relationship extraction[6, 7]. In addition to obtaining performance improvements over a strong baseline model, LCMs utilize a sparse attention mechanism which can detect context cue spans in an unsupervised fashion, lending interpretability to the model predictions. As CMED includes

[*] Corresponding author
[1] Equal contributor

the upstream tasks of Medication Detection (MD) and Event Extraction (EE), in addition to Context Classification (CC), we also develop models for these tasks and report their performance in the end-to-end setting in order to facilitate future research in this area. Our code is made publicly available at `https://github.com/jvasilakes/n2c2-track1`.

2. Related Work

We discuss prior work on context classification, cue discovery, and levitated markers in the following sections.

2.1. Context Classification

It has long been understood that linguistic context is a necessary part of clinical text mining systems. One of the earliest was NegEx, a rule-based system designed to identify negation in discharge summaries [8]. Even after the rise of machine- and deep-learning in clinical NLP, rule-based systems continued to be useful: NegEx has been extended with dependency information [9] and to languages other than English [10], NegBio [11] uses a rule-based approach to identify negation and uncertainty, and SemRep —a rule-based system for relationship extraction— has been extended to predict the factuality of biomedical events [3].

Still, deep-learning approaches, such as those based on pre-trained transformers, have achieved state-of-the-art results on negation and speculation scope detection in biomedical text [12, 13, 14, 15]. However, these models are given gold-standard cue spans either as input to the scope resolution model or as a signal for training a cue detection model. This contrasts with CMED, where events are simply given a set of context labels without any additional cue information.

2.2. Cue Discovery

Previous works on detecting context cues are generally supervised, using a manually labeled corpus [16, 17] or a seed list [18]. However, cue discovery is related to the task of rationale extraction where models are trained to produce a reason for each prediction, without any explicit supervision. For example, [19] and [20] use differentiable masks to highlight a subset of the input text as a rationale for text classification. Our proposed LCMs are similar to these methods, but instead of explicitly predicting a rationale using a separate model with a specialized loss function, LCMs produce the rationale as a by-product of prediction using a sparse attention mechanism with the standard cross-entropy loss.

2.3. Levitated Markers

Levitated markers [6] were originally developed to speed up model inference for relation extraction. Instead of inserting marker tokens into the input text around each subject and object span (cf. [21]), marker tokens are appended to the end of the input and tied to a span by a shared position ID and a directional attention mask. Follow up work [7] achieved performance gains by strategically packing together multiple levitated markers into a single training instance, in theory allowing the model to learn correlations between object spans. In essence, this encoding strategy allows a pretrained transformer to focus on multiple levitated "context" spans for a single solid-marked "target" span, without extensive augmentation of the input text. LCMs leverage this to utilize potentially many context spans for a given medication change event, pooling the context representations rather than making separate predictions for each. A comparison of our method to the previous work is given in Figure 1.

3. Materials and Methods

3.1. Dataset and Tasks

CMED contains 500 clinical notes from 296 patients and was annotated for three subtasks:

- **Medication Detection** (MD): Find all medication spans in the input text.

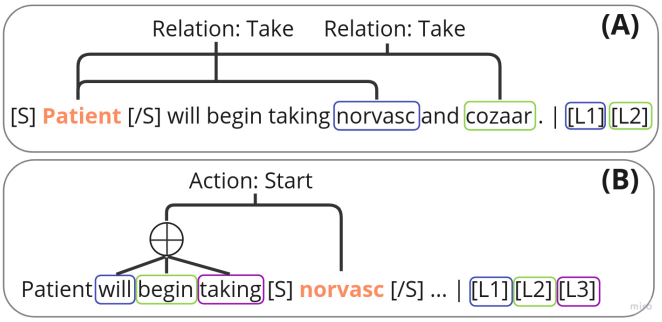

Relation: Take
Relation: Take
(A)
[S] Patient [/S] will begin taking norvasc and cozaar . | [L1] [L2]
Action: Start
(B)
Patient will begin taking [S] norvasc [/S] ... | [L1] [L2] [L3]

Figure 1: Comparison of levitated markers for their original use case of relationship extraction (A) and our task of context classification (B). In (A), a hypothetical relationship extraction task, the subject of the relation is given solid markers (`[S]`, `[/S]`) and all candidate objects are given levitated markers (e.g., `[L1]`, with colors indicating shared position IDs). A classifier then makes a prediction for each subject-object pair. In (B), the medication span is given solid markers and multiple spans are given levitated markers to act as the context. Instead of making a prediction for each pair, the levitated span representations are pooled and used to make a single prediction.

- **Medication Change Event Extraction** (EE): For each medication mention, determine if a medication change is being discussed (Disposition) or not (NoDisposition), or if it is unclear (Undetermined).
- **Context Classification** (CC): Classify each Disposition event according to 5 context dimensions, described in Table 1.

The 500 total notes are split into 350 for model training, 50 for development, and 100 for testing. Each subtask relies on a subset of labels from the previous subtask. Therefore, while the MD task encompasses 350 training notes and 6,196 medication mentions, only 340 training notes contain any medication spans for input to the EE task, and there are only 272 notes containing 1,1191 Disposition events as input to the CC task. Detailed datasets statistics are given in Appendix C.

3.2. Models

This section describes LCMs for the EE and CC tasks, as well as a span-based model for the MD task.

Context	Definition	Labels
Action	What medication change is being discussed.	Start, Stop, Increase, Decrease, OtherChange, UniqueDose, Unknown
Actor	Who initiated the change.	Physician, Patient, Unknown
Certainty	How likely the change is to occur.	Certain, Hypothetical, Conditional, Unknown
Negation	Whether the change is negated.	Negated, NotNegated
Temporality	When the change is said to occur.	Past, Present, Future, Unknown

Table 1: Context dimensions along with their definitions and possible labels.

3.2.1. Medication Detection

We treat the task of Medication Detection as a Named Entity Recognition (NER) task. Given the input text with n tokens $\{w_0, w_1, ..., w_n\}$, a pretrained language model (PLM) encoder [22] produces the hidden token representations $\{\boldsymbol{h}_0, \boldsymbol{h}_1, ..., \boldsymbol{h}_m\}$. Similarly to [23], we exhaustively extract all possible spans with a maximum length of L from the input. Each span embedding is calculated as

$$\boldsymbol{s}_{i,j} = \left[\boldsymbol{h}_i ; \frac{\Sigma_{t=i}^{j} \boldsymbol{h}_t}{j - i + 1} ; \boldsymbol{h}_j \right] \tag{1}$$

where i and j are the start and end positions of the span and $[\cdot;\cdot]$ denotes vector concatenation.

Span embeddings are then input to a binary classifier with sigmoid activation to classify the spans into entity or non-entity. Model estimation uses the binary cross entropy loss.

3.2.2. Event Extraction and Context Classification

We employ the same fundamental model for the EE and CC tasks. Input text containing the target medication mention is first preprocessed by inserting solid markers before and after the mention span. We also append levitated markers (described in Section 3.2.3) to the input, which represent task-specific context. This is then passed through a PLM encoder to obtain a hidden representation of the medication mention and context, which is input to a classification layer for prediction. Formally,

$$\boldsymbol{h}_m = \mathrm{PLM}(x_m) \tag{2}$$

$$\boldsymbol{h}_\ell = \mathrm{PLM}(\boldsymbol{\ell}) \tag{3}$$

$$\hat{y} = f([\boldsymbol{h}_m;\ c(\boldsymbol{h}_\ell)]) \tag{4}$$

where x_m is the medication mention, $\boldsymbol{\ell} = \{\ell_0, \ell_1, ..., \ell_L\}$ are the levitated markers, and $f(\cdot)$ is the classification function. The $c(\cdot)$ function pools the hidden representations of the levitated markers into a fixed-length context vector. Model estimation uses the standard cross entropy loss of the predicted class against the gold label.

The following subsection describes levitated markers and how we compute the context vector from them.

3.2.3. Levitated Context Markers

Previous work using levitated markers for relationship extraction assigned solid markers to subject spans and levitated markers to potential objects, making predictions for each subject-object pair [7]. In contrast, our LCMs pool multiple levitated markers related to a single Disposition event in order to focus the model on potentially useful subspans of the input, beyond simply the medication span representation computed from the global input context. We experiment with two methods for choosing which spans to assign levitated markers.

1. Window: Mark all spans in a $W = \frac{L}{2}$ window before and after the target medication mention, where L is the total number of levitated markers.
2. Rule-based: Mark spans according to some rule or based on external knowledge of the task. We choose the following rules for each task:
    - Event: we mark all verbs and auxiliary verbs in the input. Additionally, we experimented with adding specific types to verbs indicating change, using the semantic tagger PyMUSAS[2] (cf. Appendix B for details).
    - Action, Temporality: We mark all verbs and auxiliary verbs in the input.
    - Actor: We mark all nouns and pronouns in the input.
    - Certainty, Negation: We mark spans using lists of negation and uncertainty cues extracted from the SFU review corpus [24].

For the above rules, parts of speech are labeled using ScispaCy [25].

We propose three different definitions of the context function $c(\cdot)$ in Equation (4) to combine the levitated marker representations.

1. Max Pooling: We compute the maximum value of each dimension in the hidden representation across all levitated markers. The result is a single vector of dimension D[3].

$$c(\boldsymbol{h}_\ell) := \max\nolimits_{i\in L}\ \boldsymbol{h}_{\ell_{ij}},\ \forall j \in D \tag{5}$$

2. Softmax Attention Pooling: We introduce an attention mechanism between the medication span and each levitated marker. The learned attention weights are then used to compute a weighted average of the levitated marker representations.

$$c(\boldsymbol{h}_\ell) := \sum_{i=1}^{L} \alpha_{mi}\boldsymbol{h}_{\ell_i} \tag{6}$$

$$\boldsymbol{\alpha}_m = \boldsymbol{\rho}(e_m) = \frac{\exp(e_{mi})}{\sum_{i=1}^{L}\exp(e_{mi})},\ \forall i \in L \tag{7}$$

$$e_{mi} = a([\boldsymbol{h}_m; \boldsymbol{h}_{\ell_i}]) \tag{8}$$

where a is an alignment model that outputs a score for each medication span-marker pair[4], and ρ is the

[2] `https://ucrel.github.io/pymusas/`

[3] Max pooling consistently outperformed mean pooling in our preliminary experiments.

[4] We use a single linear layer $\hat{y} = \tanh([h_m; h_{\ell_i}]w^\top),\ w \in \mathbb{R}^{(2D)}$

softmax function which projects the raw attention scores $\boldsymbol{e}_m \in \mathbb{R}^L$ into probabilities $\boldsymbol{\alpha}_m \in \triangle^{(L-1)}$.

3. Sparse Attention Pooling: We replace the softmax projection function ρ in Equation (7) with a sparse version, Sparsegen-lin [26], which was previously used for sparse self-attention in transformer models [27].

$$\rho(e_{mi}) = \max\left\{0, \frac{e_{mi} - \tau(\boldsymbol{e}_m)}{1-\lambda}\right\} \tag{9}$$

where $\tau(\cdot)$ is a thresholding function, which ensures $\sum_{i=1}^{L} \rho(e_{mi}) = 1$ and is the result of solving an optimization problem given $\boldsymbol{e}_m$ and λ[5]. The $\lambda < 1$ hyperparameter controls how many scores in $\boldsymbol{e}_m$ become 0 in the resulting probability vector. Specifically, as $\lambda \to 1$, the result approaches a one-hot vector and as $\lambda \to -\infty$ the result approaches uniform. In theory, a sparse probability vector will force the model to select only those levitated spans that are highly correlated with the downstream task. We can then inspect these probabilities at inference to gain some interpretability of the model's predictions.

4. Results

For MD and EE, we report the average performance on the development set across 6 different cross-validation splits. Due to the small size and imbalanced class distributions of the CC data, we average results over 15 runs: 5 different cross-validation splits and 3 different random model initializations. The best performing model on the development set is then evaluated on the test set, having been trained on the standard train split for reproducibility. We report test set performance given the gold-standard input and in the end-to-end setting, i.e., given the predictions from the best model for the previous subtask. We fix model hyperparameters according to our participation in the n2c2 shared task, which are detailed in Appendix A.

4.1. Medication Detection

We experimented with two PLMs for our MD model: ClinicalBERT [28] and BioLM [29]. We evaluate the models using the strict and lenient span matching precision, recall, and F1 score. The results are given in Table 2. BioLM outperformed ClinicalBERT according to all metrics, achieving an average strict F1 score of 0.972 on the development set and 0.963 on the test set.

Dev set	Strict			Lenient		
	P	R	F1	P	R	F1
Clinical	0.972	0.961	0.966	0.979	0.968	0.974
BioLM	**0.979**	**0.965**	**0.972**	**0.983**	**0.969**	**0.976**
Test set	BioLM					
Gold	0.960	0.965	0.963	0.971	0.976	0.974

Table 2: Macro averaged (P)recision, (R)ecall, and F1 score of the medication detection task on the development set and the test set of CMED. Bolded numbers indicate the best performance on the development set. Clinical: ClinicalBERT.

4.2. Event Extraction

We experiment with LCMs using max pooling and the rule-based approach discussed in Section 3.2.3. Both attention pooling methods and the window-based levitated markers underperformed max pooling with rule-based levitated markers in preliminary experiments, so we omit these results for space reasons. Our baseline model uses solid markers surrounding the medication span only.

We report the micro- and macro-averaged results on the EE task in Table 3. While max pooled LCMs perform poorer than the baseline, adding in verb type information results in the best performance according to both the micro and macro averaged metrics. We discuss some limitations of LCMs in Section 5.2.

4.3. Context Classification

We estimate a separate model for each CC dimension, experimenting with the three context pooling functions described in Section 3.2.3. The levitated window size W and

[5] We refer the reader to Appendix A.1 of [26] for details.

Dev set	Micro			Macro		
	P	R	F1	P	R	F1
Baseline	0.930	0.928	0.929	**0.882**	0.827	0.854
LCMs	0.929	0.925	0.927	0.877	0.831	0.853
+types	**0.933**	**0.930**	**0.931**	0.875	**0.842**	**0.858**
Test set	LCMs+types					
Gold	0.926	0.927	0.926	0.842	0.833	0.837
E2E	0.896	0.907	0.901	0.805	0.822	0.813

Table 3: Micro and Macro averaged (P)recision, (R)ecall, and F1 score of event detection on the development set of CMED. **Baseline**: No levitated markers; solid markers surrounding the medication span only. **LCMs**: Solid markers and max pooled LCMs. **+types**: Typed levitated markers according to the semantic tags of the verbs. The best numbers in each column are in **bold**. The typed model was evaluated on the test set using gold standard medication spans (Gold) and predictions from the best MD model (E2E). For the end-to-end results, we report the performance for lenient span matching.

Sparsegen-lin λ coefficient were tuned on the development set. We also report performance of a baseline model with no levitated markers, where predictions are made given only the solid markers surrounding the medication span.

Classification results for each model and each context dimension are given in Table 5. As already shown for EE in Table 3, our baseline is quite strong, outperforming many of our models. Still, with the exception of the Action dimension, LCMs with windowed sparse attention pooling outperform the baseline with solid markers only. However, there is a large performance drop between development and test sets across dimensions. We discuss this disparity in more detail in Section 5.1. While using the rules described in Section 3.2.3 to choose levitated spans outperforms the baseline in some cases (e.g., Temporality and Negation), its performance is inconsistent. We provide a potential explanation for this in Section 5.

4.4. Cue Detection

We performed a manual evaluation of the cues detected by the sparse attention pooling mechanism by inspecting the spans with the highest attention weight. We note that, while softmax pooling outperformed sparse pooling for Action dimension, softmax tended to produce approximately uniform weights across spans, meaning it was not useful for cue detection.

An annotator (JV) marked a random, stratified subset of 50 cue predictions on the development set from each dimension as either "Relevant" or "Irrelevant" to the predicted label. Due to the infrequency of "Negated" examples, Negation cue predictions were sourced from both the development and train sets. The full annotation guidelines are available alongside our code. We report the accuracy for each dimension as the percentage of "Relevant" cues.

We performed this evaluation for each context dimension with the following exceptions: we excluded the Actor dimension as nearly all cues were either implicit or in the document-level context (cf. the example in Figure 2); for Certainty we excluded the "Certain" label, as there are no cues for these examples; likewise, we only evaluated "Negated" instances.

Table 4 shows the percentage of Relevant cues detected by the sparse attention pooling mechanism for each context dimension. We also provide examples of Relevant cues detected for each context dimension in Figure 2. Additional Relevant and Irrelevant cues are given in Figure E.4.

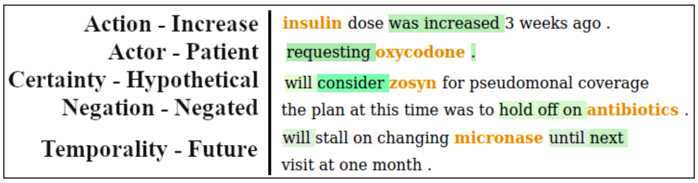
We see that the cues detected by the sparse attention mechanism are often relevant to the predicted label, at around 60% for Action, Negation, and Temporality. For Action, the detected cues are relevant verbs such as "increase", "add", or "reduce". For Negation, detected cues include both syntactic (e.g., "not") and lexical (e.g., "hold

Dimension	Accuracy	Dimension	Accuracy
Action	0.653	Negation	0.600
Certainty	0.441	Temporality	0.563

Table 4: Accuracy of the cues detected by the sparse attention pooling mechanism, computed as the percentage of Relevant cues.

Dev set	Action			Actor			Certainty			Negation			Temporality		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baseline	0.837	0.786	**0.811**	0.766	0.716	0.740	0.686	0.673	0.679	0.781	0.657	0.713	0.792	0.773	0.783
Maxpool															
window	0.834	0.785	0.809	0.751	0.713	0.731	0.679	0.668	0.673	0.730	0.677	0.703	**0.801**	0.797	0.799
rules	0.825	0.784	0.804	0.722	0.730	0.725	0.661	0.652	0.656	0.789	0.679	0.730	0.795	0.792	0.793
Softmax															
window	0.836	**0.788**	**0.811**	0.758	0.706	0.731	0.682	0.667	0.674	0.796	0.679	0.733	0.769	0.785	0.777
rules	**0.838**	0.781	0.809	0.745	**0.732**	0.738	0.682	0.661	0.671	0.766	0.632	0.693	0.783	0.780	0.782
Sparse															
window	0.835	0.780	0.806	**0.799**	0.709	**0.751**	**0.707**	**0.676**	**0.691**	0.774	**0.697**	**0.734**	0.769	**0.809**	**0.803**
rules	0.833	0.784	0.808	0.733	0.717	0.725	0.678	0.660	0.669	**0.801**	0.673	0.732	0.768	0.787	0.777
Test set	Softmax-window			Sparse-window			Sparse-window			Sparse-window			Sparse-window		
Gold	0.881	0.744	0.793	0.559	0.549	0.554	0.583	0.626	0.591	0.491	0.500	0.495	0.590	0.592	0.591
E2E	0.751	0.659	0.689	0.537	0.441	0.472	0.532	0.566	0.547	0.392	0.412	0.402	0.506	0.494	0.498

Table 5: Macro averaged (P)recision, (R)ecall, and F1 score of context classification on CMED for each pooling method and context dimension. The best score for each metric is in **bold**. The model for each dimension with the best F1 score on the development set was evaluated on the test set given gold standard event input (Gold) and predictions from the best EE model (E2E). **Baseline**: No levitated markers; solid markers surrounding the event span only. **Maxpool**: Max pooling of the levitated marker representations. **Softmax**: Attention pooling with softmax projection. **Sparse**: Attention pooling with Sparsegen-lin projection.



Action - Increase	insulin dose was increased 3 weeks ago .
Actor - Patient	requesting oxycodone .
Certainty - Hypothetical	will consider zosyn for pseudomonal coverage
Negation - Negated	the plan at this time was to hold off on antibiotics .
Temporality - Future	will stall on changing micronase until next visit at one month .

Figure 2: Examples of Relevant cues from each context dimension. The medication is in orange and cues are highlighted in green proportional to their attention weight.

off", "declines"). For Temporality, the cues include both auxiliary verbs (e.g., "will") and verbs of a relevant tense (e.g., "started" for Past). Certainty is an exception, with only 44% relevant cues. While cues such as "consider" are often correctly identified, others such as "would" and "suspect" can be missed (cf. Figure E.4d).

5. Discussion

5.1. Error Analysis

As mentioned in Section 4.3, there is a large performance drop between development and test sets for all context dimensions except Action. We here provide an explanation of this performance difference by way of a brief error analysis.

For Actor, Certainty, and Temporality the performance drop is due almost exclusively to poor performance on the "Unknown" label. This label is very infrequent in the dataset, making it very difficult to learn generalizable representations (cf. Table C.9 for dataset statistics). For example, there are six "Unknown" Certainty examples in the test set, only two in the train set, and none in the development set. Additionally, few test examples for a given label means that a single incorrect prediction has a large effect on the macro averaged metrics. In Appendix D, we provide per-label performance on the development and test sets for each context dimension, which show that the performance drop on these dimensions is due exclusively to the "Unknown" label.

The performance discrepancy for Negation is due to poor test performance on the "Negated" label, which is quite infrequent. A review of Negated instances across the splits revealed a large variety of expressions of negation including lexical (e.g., "Coumadin was *deferred*") and requiring anaphora resolution (e.g., "We dicussed narcotics. We *elected against* these medications."). Few examples and a variety of expressions again means learning generalizable representations is difficult.

5.2. Limitations and Future Work

This study has the following limitations: The small size of the development and test sets for the CC subtask means that it is difficult to be certain regarding performance differences between methods. We addressed this by training and evaluating our models on multiple cross-validation splits and with multiple random initializations, but the confidence intervals for each metric are quite wide (around 20% across dimensions). As discussed in Section 5.1, the small dataset size also hinders the models' ability to generalize to the test examples. Future work should therefore expand the amount of data available for both training and evaluation, especially regarding the less frequent labels.

We also found that our rule-based levitated markers could be too limiting. For example, while marking verbs our model missed clinical shorthand cues such as "d/c" ("discontinued") and our negation cue list was missing some lexical cues (e.g., "hold", "deny") that were prevalent in the data. Missed verbs were therefore excluded from the subsequent semantic tagging for the EE task and were never assigned the relevant type. Future work will be to develop a hybrid approach in which rules guide the model without overly restricting it.

6. Conclusion

We proposed Levitated Context Markers (LCMs), a novel method for event extraction and context classification in clinical text. LCMs utilize shared position IDs and a directional attention mask to allow pretrained transformers to learn from the global context as well as focus on task-specific subspans without extensive markup of the input. Experiments on the Contextualized Medication Extraction Dataset show that LCMs outperform a strong transformer baseline model on event extraction and context classification. Additionally, LCMs use a sparse attention mechanism which is able to detect relevant cue spans —e.g., negation triggers, modals, and verbs of change— in an unsupervised fashion, adding interpretability to the model predictions.

Acknowledgements

This work was supported by a collaboration between the University of Manchester and the Artificial Intelligence Research Center (BBSRC Japan Partnering Award BB/P025684/1), the Alan Turing Institute, as well as the University of Manchester President's Doctoral Scholar award, the Engineering and Physical Sciences Research Council CASE award, and Atypon Systems Limited, UK.

Appendix A. Implementation Details

All models were implemented using PyTorch [30]. Model training and evaluation was performed on a NVIDIA Tesla A100 with 40GB of VRAM. Hyperparameter settings for each model are detailed in tables A.6, A.7, and A.8. Hyperparameters for the medication detection models were tuned using Optuna [31].

Batch size	16
Epochs	10
Learning rate	{1e-5, 3e-4}
Max span length	{6,8,10,12,14}
Max sentence length	{128, 256}
Number of trials	30

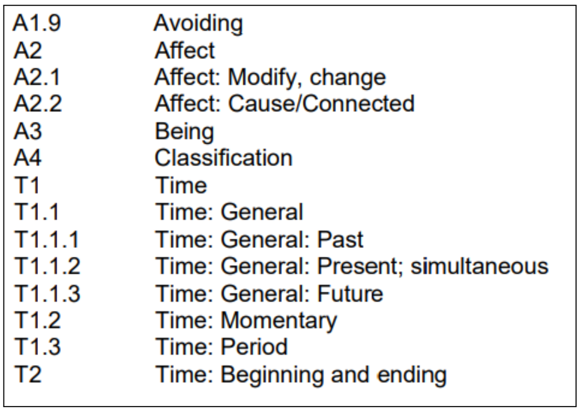
Table A.6: Hyperparameter settings for the medication detection models.

PLM	BlueBERT
Max sequence length	300
Batch size	32
Epochs	10
Learning rate	4e-5
Auxiliary tasks	Action
Pooling method	Max pooling

Table A.7: Hyperparameter settings for the Event Extraction models.

Appendix B. Semantic Tagging Details

PyMUSAS provides a hierarchical tag-set[6] along 21 major discourse fields that expand into 232 fine-grained category labels. Each item is tagged with one or more semantic labels and there is also category *Z99* for unmatched spans. We choose categories *A2.1* and *T2*, since they correspond to the verbs “increase”, “decrease” and “start”, “stop” respectively, and assigned levitated markers different tokens for verbs belonging to these categories.



A1.9	Avoiding
A2	Affect
A2.1	Affect: Modify, change
A2.2	Affect: Cause/Connected
A3	Being
A4	Classification
T1	Time
T1.1	Time: General
T1.1.1	Time: General: Past
T1.1.2	Time: General: Present; simultaneous
T1.1.3	Time: General: Future
T1.2	Time: Momentary
T1.3	Time: Period
T2	Time: Beginning and ending

Figure B.3: Part of PyMUSAS fine-grained categories.

[6] https://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf

	Action	Actor	Certainty	Negation	Temporality
PLM	BlueBERT	BERT-base	ClinicalBERT	ClinicalBERT	ClinicalBERT
Max sequence length	300	300	300	300	300
Batch size	32	32	32	32	32
Epochs	20	10	20	20	20
Learning rate	3e-5	3e-5	3e-5	3e-5	3e-5
Auxiliary data	i2b2 2009	-	i2b2 2009	i2b2 2009	i2b2 2009
Auxiliary tasks	-	-	-	Action	-
W	10	5	10	10	5
λ	0.5	0.5	0.5	0.2	0.5

Table A.8: Hyperparameter settings for the Context Classification models. W: window size parameter for the Levitated Context Markers. λ: sparsity coefficient for the Sparsegen-lin projection function.

Appendix C. Dataset Statistics

We provide per-label counts for the Event and Context Classificaton tasks in Table C.9.

Task	Label	Train	Dev	Test
Event	NoDisposition	4535	725	1326
	Disposition	1191	221	335
	Unknown	470	87	122
Action	Start	471	97	131
	Stop	280	60	67
	UniqueDose	263	22	88
	Increase	106	23	22
	Decrease	41	13	13
	Unknown	29	6	14
	OtherChange	1	0	0
Actor	Physician	1084	194	311
	Patient	89	17	17
	Unknown	18	10	7

Task	Label	Train	Dev	Test
Negation	NotNegated	1163	217	329
	Negated	28	4	6
Certainty	Certain	1001	175	281
	Hypothetical	105	29	33
	Conditional	83	17	15
	Unknown	2	0	6
Temporality	Past	613	131	173
	Present	440	54	132
	Future	112	33	29
	Unknown	26	3	1

Table C.9: Dataset statistics for the train, development, and test sets for the Event Extraction and Context Classification tasks.

Appendix D. Additional Results

Complementary to Table 5 in the main text, Table D.10 provides per-label precision, recall, and F1 scores for each context dimension on the standard development set and the test set.

Action	Dev				Test			
	P	R	F1	N	P	R	F1	N
Decrease	0.727	0.615	0.667	13	0.800	0.615	0.696	13
Increase	1.000	0.826	0.905	23	0.882	0.682	0.769	22
Start	0.802	0.918	0.856	97	0.829	0.962	0.890	131
Stop	0.902	0.767	0.829	60	0.821	0.821	0.821	67
UniqueDose	0.826	0.864	0.844	22	0.951	0.886	0.918	88
Unknown	0.833	0.833	0.833	6	1.000	0.500	0.667	14

Actor	Dev				Test			
	P	R	F1	N	P	R	F1	N
Patient	0.727	0.471	0.571	17	0.562	0.529	0.545	17
Physician	0.936	0.979	0.957	194	0.971	0.974	0.973	311
Unknown	0.571	0.400	0.471	10	0.143	0.143	0.143	7

Certainty	Dev				Test			
	P	R	F1	N	P	R	F1	N
Certain	0.919	0.971	0.944	175	0.948	0.964	0.956	281
Conditional	0.769	0.588	0.667	17	0.583	0.933	0.718	15
Hypothetical	0.826	0.655	0.731	29	0.800	0.606	0.690	33
Unknown	-	-	-	0	0.000	0.000	0.000	6

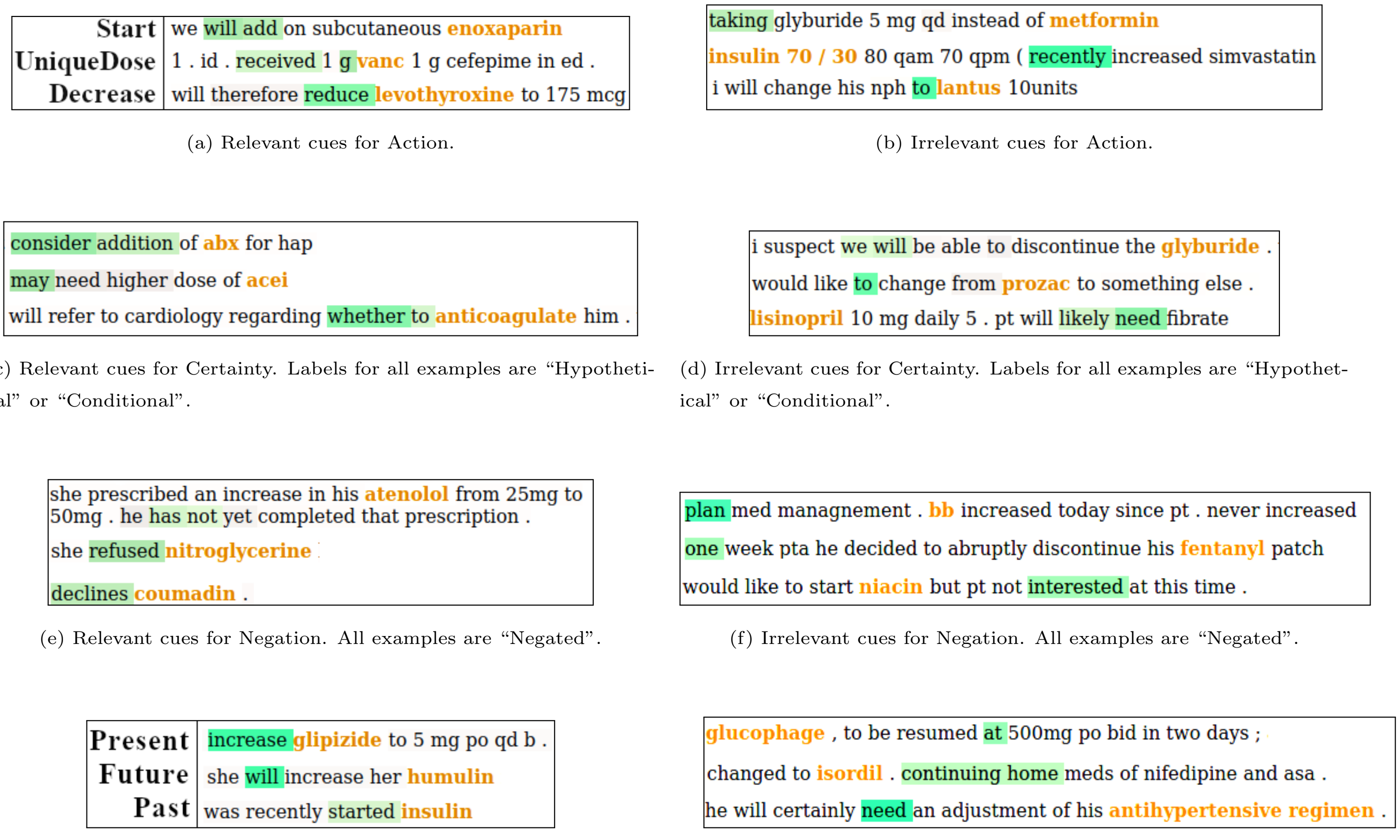
Negation	Dev				Test			
	P	R	F1	N	P	R	F1	N
Negated	1.000	0.250	0.400	4	0.000	0.000	0.000	6
NotNegated	0.986	1.000	0.993	217	0.982	1.000	0.991	329

Temporality	Dev				Test			
	P	R	F1	N	P	R	F1	N
Future	0.786	0.667	0.721	33	0.621	0.621	0.621	29
Past	0.944	0.893	0.918	131	0.940	0.908	0.924	173
Present	0.662	0.833	0.738	54	0.799	0.841	0.819	132
Unknown	1.000	0.333	0.500	3	0.000	0.000	0.000	1

Table D.10: Per-label (P)recision, (R)ecall, F1 score, and (N)umber of examples for each context dimension on the standard development split and the test split.

Appendix E. Cue Detection Examples

Figure E.4 supplements Figure 2 in the main text with additional “Relevant” and “Irrelevant” detected cues for each context dimension besides Action.



Start	we will add on subcutaneous enoxaparin
UniqueDose	1 . id . received 1 g vanc 1 g cefepime in ed .
Decrease	will therefore reduce levothyroxine to 175 mcg

(a) Relevant cues for Action.

taking glyburide 5 mg qd instead of metformin
insulin 70 / 30 80 qam 70 qpm (recently increased simvastatin
i will change his nph to lantus 10units

(b) Irrelevant cues for Action.

consider addition of abx for hap
may need higher dose of acei
will refer to cardiology regarding whether to anticoagulate him .

(c) Relevant cues for Certainty. Labels for all examples are “Hypothetical” or “Conditional”.

i suspect we will be able to discontinue the glyburide .
would like to change from prozac to something else .
lisinopril 10 mg daily 5 . pt will likely need fibrate

(d) Irrelevant cues for Certainty. Labels for all examples are “Hypothetical” or “Conditional”.

she prescribed an increase in his atenolol from 25mg to 50mg . he has not yet completed that prescription .
she refused nitroglycerine
declines coumadin .

(e) Relevant cues for Negation. All examples are “Negated”.

plan med managnement . bb increased today since pt . never increased
one week pta he decided to abruptly discontinue his fentanyl patch
would like to start niacin but pt not interested at this time .

(f) Irrelevant cues for Negation. All examples are “Negated”.

Present	increase glipizide to 5 mg po qd b .
Future	she will increase her humulin
Past	was recently started insulin

(g) Relevant cues for Temporality.

glucophage , to be resumed at 500mg po bid in two days ;
changed to isordil . continuing home meds of nifedipine and asa .
he will certainly need an adjustment of his antihypertensive regimen .

(h) Irrelevant cues for Temporality.

Figure E.4: Example cues detected by LCMs with sparse attention pooling on each context dimension. Gold labels are provided where necessary to aid interpretation of the cues.

References

[1] A. Turchin, M. Shubina, E. Breydo, M. L. Pendergrass, J. S. Einbinder, Comparison of information content of structured and narrative text data sources on the example of medication intensification, Journal of the American Medical Informatics Association 16 (3) (2009) 362–370. doi:10.1197/jamia.M2777.

[2] P. Bose, S. Srinivasan, W. C. Sleeman, J. Palta, R. Kapoor, P. Ghosh, A survey on recent named entity recognition and relationship extraction techniques on clinical texts, Applied Sciences 11 (1818) (2021) 8319. doi:10.3390/app11188319.

[3] H. Kilicoglu, G. Rosemblat, T. C. Rindflesch, Assigning factuality values to semantic relations extracted from biomedical research literature, PLOS ONE 12 (7) (2017) e0179926. doi:10.1371/journal.pone.0179926.

[4] A. L. Olex, B. T. McInnes, Review of temporal reasoning in the clinical domain for timeline extraction: Where we are and where we need to be, Journal of Biomedical Informatics 118 (2021) 103784. doi:10.1016/j.jbi.2021.103784.

[5] D. Mahajan, J. J. Liang, C.-H. Tsou, Toward understanding clinical context of medication change events in clinical narratives, AMIA Annual Symposium Proceedings 2021 (2022) 833–842.

[6] Z. Zhong, D. Chen, A frustratingly easy approach for entity and relation extraction, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, p. 50–61. doi:10.18653/v1/2021.naacl-main.5.
URL https://aclanthology.org/2021.naacl-main.5

[7] D. Ye, Y. Lin, P. Li, M. Sun, Packed levitated marker for entity and relation extraction, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 4904–4917. doi:10.18653/v1/2022.acl-long.337.

[8] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, B. G. Buchanan, A simple algorithm for identifying negated findings and diseases in discharge summaries, Journal of Biomedical Informatics 34 (5) (2001) 301–310. doi:10.1006/jbin.2001.1029.

[9] S. Sohn, S. Wu, C. G. Chute, Dependency parser-based negation detection in clinical narratives, AMIA Summits on Translational Science Proceedings 2012 (2012) 1–8.

[10] W. W. Chapman, D. Hilert, S. Velupillai, M. Kvist, M. Skeppstedt, B. E. Chapman, M. Conway, M. Tharp, D. L. Mowery, L. Deleger, Extending the negex lexicon for multiple languages, Studies in health technology and informatics 192 (2013) 677–681.

[11] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, Z. Lu, Negbio: a high-performance tool for negation and uncertainty detection in radiology reports, AMIA Summits on Translational Science Proceedings 2018 (2018) 188–196.

[12] A. Khandelwal, B. K. Britto, Multitask learning of negation and speculation using transformers, in: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, Association for Computational Linguistics, Online, 2020, p. 79–87. doi:10.18653/v1/2020.louhi-1.9.
URL https://aclanthology.org/2020.louhi-1.9

[13] A. Khandelwal, S. Sawant, Negbert: A transfer learning approach for negation detection and scope resolution, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, p. 5739–5748.
URL https://www.aclweb.org/anthology/2020.lrec-1.704

[14] M. Hartmann, A. Søgaard, Multilingual negation scope resolution for clinical text, in: Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, Association for Computational Linguistics, online, 2021, p. 7–18.
URL https://aclanthology.org/2021.louhi-1.2

[15] P. Tiwary, A. Madhubalan, A. Gautam, No means ‘no’: a non-improper modeling approach, with embedded speculative context, Bioinformatics 38 (20) (2022) 4790–4796. doi:10.1093/bioinformatics/btac593.

[16] G. Szarvas, V. Vincze, R. Farkas, J. Csirik, The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts, in: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, BioNLP ’08, Association for Computational Linguistics, USA, 2008, p. 38–45.

[17] P. Thompson, R. Nawaz, J. McNaught, S. Ananiadou, Enriching a biomedical event corpus with meta-knowledge annotation, BMC Bioinformatics 12 (1) (2011) 393. doi:10.1186/1471-2105-12-393.

[18] C. Chen, M. Song, G. E. Heo, A scalable and adaptive method for finding semantically equivalent cue words of uncertainty, Journal of Informetrics 12 (1) (2018) 158–180. doi:10.1016/j.joi.2017.12.004.

[19] T. Lei, R. Barzilay, T. Jaakkola, Rationalizing neural predictions, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, p. 107–117. doi:10.18653/v1/D16-1011.
URL https://aclanthology.org/D16-1011

[20] J. Bastings, W. Aziz, I. Titov, Interpretable neural predictions with differentiable binary variables, in: Proceedings of the 57th

Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, p. 2963–2977. doi:10.18653/v1/P19-1284.
URL https://aclanthology.org/P19-1284

[21] L. Baldini Soares, N. FitzGerald, J. Ling, T. Kwiatkowski, Matching the blanks: Distributional similarity for relation learning, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, p. 2895–2905. doi:10.18653/v1/P19-1279.
URL https://www.aclweb.org/anthology/P19-1279

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017, pp. 6000–6010.

[23] M. G. Sohrab, M. Miwa, Deep exhaustive model for nested named entity recognition, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2843–2849. doi:10.18653/v1/D18-1309.

[24] N. Konstantinova, S. C. de Sousa, N. P. Cruz, M. J. Maña, M. Taboada, R. Mitkov, A review corpus annotated for negation, speculation and their scope, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, p. 3190–3195.
URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/533_Paper.pdf

[25] M. Neumann, D. King, I. Beltagy, W. Ammar, ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019, pp. 319–327. arXiv:arXiv:1902.07669, doi:10.18653/v1/W19-5034.
URL https://www.aclweb.org/anthology/W19-5034

[26] A. Laha, S. A. Chemmengath, P. Agrawal, M. Khapra, K. Sankaranarayanan, H. G. Ramaswamy, On controllable sparse alternatives to softmax, in: Advances in Neural Information Processing Systems, Vol. 31, Curran Associates, Inc., 2018, pp. 6423—-6433.

[27] B. Cui, Y. Li, M. Chen, Z. Zhang, Fine-tune bert with sparse self-attention mechanism, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, p. 3548–3553. doi:10.18653/v1/D19-1361.
URL https://aclanthology.org/D19-1361

[28] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 72–78. doi:10.18653/v1/W19-1909.

[29] P. Lewis, M. Ott, J. Du, V. Stoyanov, Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art, in: Proceedings of the 3rd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Online, 2020, pp. 146–157. doi:10.18653/v1/2020.clinicalnlp-1.17.

[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 32, Curran Associates, Inc., 2019, pp. 8026—-8037.
URL https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

[31] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 2623–2631. doi:10.1145/3292500.3330701.

Contextualized Medication Event Extraction with Levitated Markers

Credit Author Statement

Jake Vasilakes: Investigation, Methodology, Software, and Validation for the context classification task, Writing - Original Draft.
Panagiotis Georgiadis: Investigation, Methodology, Software, and Validation for the event extraction task, Writing - Original Draft.
Nhung T.H. Nguyen: Investigation, Methodology, Software, and Validation for the medication detection task, Writing - Review & Editing.
Makoto Miwa: Funding acquisition, Supervision, Writing - Review & Editing.
Sophia Ananiadou: Funding acquisition, Supervision.