This before That: Causal Precedence in the Biomedical Domain

Gus Hahn-Powell Dane Bell Marco A. Valenzuela-Escárcega Mihai Surdeanu

University of Arizona
Tucson, AZ 85721, USA
hahnpowell@email.arizona.edu

Abstract

Causal precedence between biochemical interactions is crucial in the biomedical domain, because it transforms collections of individual interactions, e.g., bindings and phosphorylations, into the causal mechanisms needed to inform meaningful search and inference. Here, we analyze *causal* precedence in the biomedical domain as distinct from open-domain, *temporal* precedence. First, we describe a novel, hand-annotated text corpus of causal precedence in the biomedical domain. Second, we use this corpus to investigate a battery of models of precedence, covering rule-based, feature-based, and latent representation models. The highest-performing individual model achieved a micro F1 of 43 points, approaching the best performers on the simpler temporal-only precedence tasks. Feature-based and latent representation models each outperform the rule-based models, but their performance is complementary to one another. We apply a sieve-based architecture to capitalize on this lack of overlap, achieving a micro F1 score of 46 points.

1 Introduction

In the biomedical domain, an enormous amount of information about protein, gene, and drug interactions appears in the form of natural language across millions of academic papers. There is a tremendous ongoing effort (Nédellec et al., 2013; Kim et al., 2012; Kim et al., 2009) to extract individual chemical interactions from these texts, but these interactions are only isolated fragments of larger causal mechanisms such as protein signaling pathways. Nowhere, however, including any database, is the complete mechanism described in a form that lends itself to causal search or inference. The absence of such a database is not for lack of trying; Pathway Commons (Cerami et al., 2011) aims to address the need, but its authors estimate it currently covers 1% of the literature due to the high cost of annotation[1]. This issue only grows more pressing with the yearly growth in biomedical publishing, which presents an otherwise insurmountable challenge for biomedical researchers to query and interpret.

The Big Mechanism program (Cohen, 2015) aims to construct exactly such large-scale mechanistic information by reading and assembling protein signaling pathways that are relevant for cancer, and exploit them to generate novel explanatory and treatment hypotheses. Although prior work (Chambers et al., 2014; Mirza, 2016) has addressed the challenging area of temporal precedence in the open domain, the biomedical domain presents very different data and, consequently, requires novel techniques. Precedence in mechanistic biology is *causal* rather than *temporal*. Though event temporality is crucial to understanding electronic health records for individual patients (Bethard et al., 2015; Bethard et al., 2016), its contribution to the understanding of biomolecular reactions is less clear as these events and processes may repeat in extremely short cycles, continue without end, or overlap in time. At any level of abstraction, causal precedence encodes mechanistic information and facilitates inference over spotty evidence. For the purpose of this work, *precedence* is defined for two events, A and B, as

> A precedes B if and only if the output of A is necessary for the successful execution of B.[2]

[1] Personal communication.

[2] See the "precedes" examples in Table 1.

Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pages 146–155,
Berlin, Germany, August 12, 2016. ©2016 Association for Computational Linguistics

Very little annotated data exists for causal precedence, especially efforts focusing on signaling pathways. BioCause (Mihăilă et al., 2013), for instance, is centered on connections between claims and evidence and contains only 51 annotated examples of causal precedence[3]. Our work[4] offers three contributions in aid of automatically extracting causal ordering in biomedical text. First, we provide and describe a dataset of real text examples, manually annotated for causal precedence. Second, we analyze the efficacy of a battery of different models in automatically determining precedence, built on top of the Reach automatic reading system (Valenzuela-Escárcega et al., 2015a; Valenzuela-Escárcega et al., 2015c) and measured against this novel corpus. In particular, we investigate three classes of models: (a) deterministic rule-based models inspired by the precedence sieves proposed by Chambers et al. (2014), (b) feature-based models, and (c) models that rely on latent representations such as long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997). Our analysis indicates that while independently the top-performing model achieves a micro F1 of 43, these models are largely complementary with a combined recall of 58 points. Lastly, we conduct an error analysis of these models to motivate and inform future research.

2 A Corpus for Causal Precedence in the Biomedical Domain

Our corpus annotates several types of relations between mentions of biochemical interactions. Following common terminology promoted by the BioNLP shared tasks, we will interchangeably use "events" to refer to these interactions. To generate candidate events for our planned annotations, we ran the Reach event extraction system (Valenzuela-Escárcega et al., 2015a; Valenzuela-Escárcega et al., 2015c) over the full text[5] of 500 biomedical papers taken from the Open Access subset of PubMed[6]. The events extracted by Reach are biochemical events of two types: simple events such as phosphorylation that modify one or more entities (typically proteins), and nested events (regulations) that have other events as arguments.

Relation	*Example*
E1 precedes E2	A is phosphorylated by B. Following its phosphorylation, A binds with C.
E2 precedes E1	A is phosphorylated by B. Prior to its phosphorylation, A binds with D.
Equivalent	The phosphorylation of A by B. A is phosphorylated by B.
E1 specifies E2	A is phosphorylated by B at Site 123. A is phosphorylated by B.
E2 specifies E1	A is phosphorylated by B. A is phosphorylated by B at Site 123.
Other	B does not regulate C when C is bound to A.
None	A phosphorylates B. A ubiquitinates C.

Table 1: The seven inter-event relation labels annotated in the corpus. The "precedes" labels are causal. Subsumption is captured with the "specifies" labels.

To improve the likelihood of finding pairs of events with a relevant link, we filtered event pairs by imposing the following requirements for inclusion in the corpus:

1. *Event pairs must share at least one participant.* This constraint is based on the observation that interactions that share participants are more likely to be connected.

2. *Event pairs must be within 1 sentence of each other.* Similarly, discourse proximity increases the likelihood of two events being related.

3. *Event pairs must not share the same type.* This helps to maximize the diversity of the dataset.

4. *Event pairs must not already be contained in an extracted Regulation event.* For example, we did not annotate the relation between the binding and the phosphorylation events in "The binding of X and Y is inhibited

[3] These are marked in the BioCause corpus as `Causality` events with `Cause` and `Effect` arguments. The remaining 800 annotations are claim-evidence relations.

[4] The corpus, tools, and system introduced in this work are publicly available at `https://github.com/myedibleenso/this-before-that`

[5] We chose to ignore the "references", "materials", and "methods" sections, which generally do not contain mechanistic information.

[6] `http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`

by X phosphorylation", because it is already captured by most state-of-the-art biomedical event extraction systems.

After applying these constraints, only 1700 event pairs remained. In order to rapidly annotate the event pairs, we developed a browser-based annotation UI that is completely client-side (see Figure 3). Using this tool, we annotated 1000 event pairs for this work; 84 of these were discarded due to severe extraction errors. The annotations include the event spans, event triggers (i.e., the verbal or nominal predicates that indicate the type of interaction such as "binding" or "phosphorylated"), source document, minimal sentential span encompassing both event mentions, and whether or not the event pair involves coreference for either the event trigger or the event participants. For events requiring coreference resolution, we expanded the encompassing span of text to also capture the antecedent. Note that domain-specific coreference resolution is a component of the event extraction system used here (Bell et al., 2016).

When describing the relations between these event pairs, we refer to the event that occurs first in text as Event 1 (E1) and the event that follows as Event 2 (E2). Each (E1, E2) pair was assigned one of seven labels: "E1 precedes E2", "E2 precedes E1", "Equivalent", "E1 specifies E2", "E2 specifies E1", "Other", or "None". Table 1 provides examples for each of these labels. We converged on these labels because they are fundamental to the assembly of causal mechanisms from a collection of events. Collectively, the seven labels address three important assembly tasks: *equivalence*, i.e., understanding that two event mentions discuss the same event, *subsumption*, i.e., the two mentions discuss the same event, but one is more specific than the other, and, most importantly, *causal precedence*, the identification of which is the focus of this work. During the annotation process, we came across examples of other relevant phenomena. We grouped these instances under the label "Other" and leave their analysis for future work.

Though simplified, the examples in Table 3 illustrate that this is a complex task sensitive to linguistic evidence. For example, the direction of the precedence relations in the first two rows in the table changes based on a single word in the context ("prior" vs. "following").

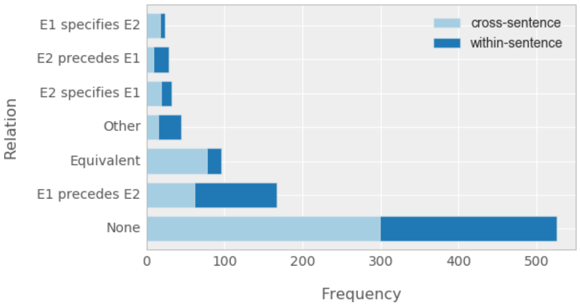


Figure 1: The distribution of assembly relation labels both within and across sentences.

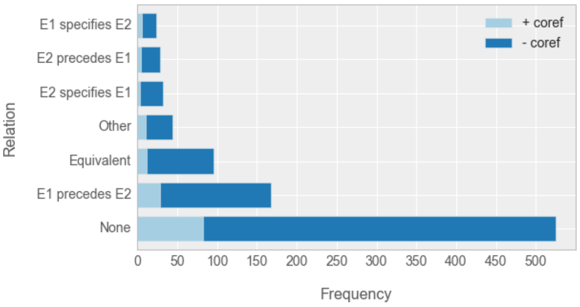


Figure 2: The distribution of event pairs involving coreference across assembly relations.

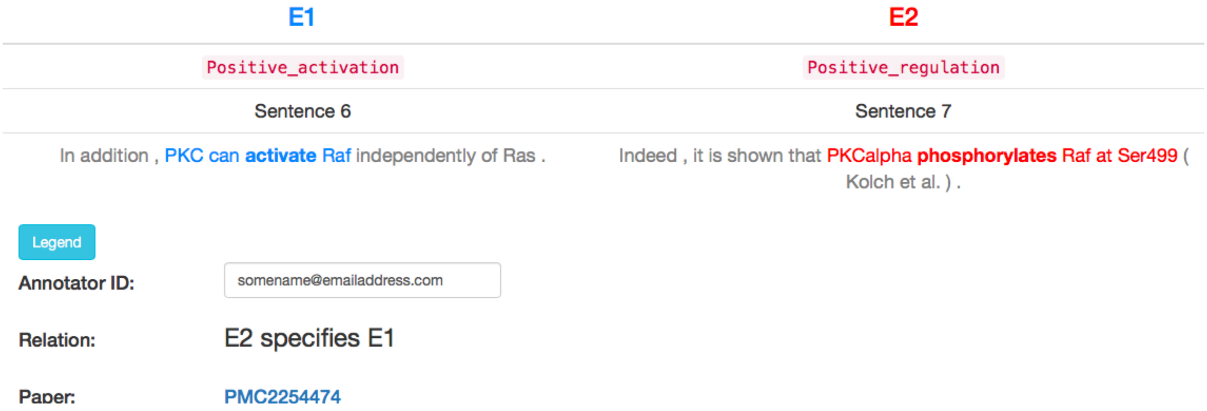
In terms of the distribution of relations, causal precedence pairs appear more frequently within the same sentence, while cases of the subsumption ("specifies") and equivalence relations are far more common across sentences (see Figure 1). Coreference is involved in 10–15% of the instances for each relation label (see Figure 2).

The annotation process was performed by two linguists familiar with the biomedical domain. To minimize errors, the annotation task was initially performed together at the same workstation.[7] On a randomly selected sample of 100 event pairs, the two annotators had a Cohen's kappa score (Cohen, 1960) of 0.82, indicating "almost perfect" agreement for the *precedes* labels (Landis and Koch, 1977).

3 Models of Causal Precedence

We have developed both deterministic, interpretable models and automatic, machine-learning models for detecting causal precedence in our dataset. Importantly, the models covered in this work focus solely on causal precedence, which is the most complex relation annotated in the dataset previously introduced. Thus, for all experiments discussed here, we reduce these annotations to three labels: "E1 precedes E2", "E2 precedes E1", and `Nil`, which covers all the other labels in the

[7] Similar to pair programming.



E1	E2
Positive_activation	Positive_regulation
Sentence 6	Sentence 7
In addition , PKC can activate Raf independently of Ras .	Indeed , it is shown that PKCalpha phosphorylates Raf at Ser499 (Kolch et al.) .

Annotator ID: somename@emailaddress.com

Relation: E2 specifies E1

Paper: PMC2254474

Figure 3: Browser-based tool for annotating assembly relations in text. An annotation instance consists of a pair of event mentions. The annotator assigns a label to each pair of events using the number keys and navigates from annotation to annotation using the arrow keys. E1 refers to the event in the pair that appears first in the text. The event span is formatted to stand out from the surrounding text. The "Paper" field provides the annotator with easy access to the full text of the source document for the current annotation instance. Annotations can be exported to JSON and reloaded via a local storage cache or through file upload.

corpus.

Model	*Rules*
Intra-sentence	29
Inter-sentence	5
Reichenbach	8

Table 2: Few rules defined each deterministic model of precedence compared with the number of features for the machine learning models.

3.1 Deterministic Models

The deterministic models are defined by a small number of hand-written rules using the Odin event extraction framework (Valenzuela-Escárcega et al., 2015b). The number of rules for each model is shown in Table 2, and sharply contrast with the 92,711 features introduced later (Table 3) that are used by our machine-learning models. In order to avoid overfitting, all of the deterministic models were created without reference to the annotation corpus, using general linguistic expertise and domain knowledge.

Intra-sentence ordering Within sentences, syntactic regularities can be exploited to cover a large variety of grammatical constructions indicating precedence relations. Rules defined over dependency parses (De Marneffe and Manning, 2008) capture precedence in sentences like those in (1) and (2) as well as many others.

(1) [The RBD of PI3KC2B binds HRAS]$_{\text{after}}$, when [HRAS is not bound to GTP]$_{\text{before}}$

(2) [The ubiquitination of A]$_{\text{before}}$ is followed by [the phosphorylation of B]$_{\text{after}}$

Other phrases captured include: "precedes", "due to", "leads to", "results in", etc.

Inter-sentence ordering Although syntax operates over single sentences, cross-sentence time expressions can indicate ordering, as shown in Examples (3) and (4). We exploit these regularities as well by checking for sentence-initial word combinations.

(3) [A is phosphorylated by B]$_{\text{before}}$. As a downstream effect, [C is . . .]$_{\text{after}}$

(4) [A is phosphorylated by B]$_{\text{before}}$. [C is then . . .]$_{\text{after}}$

Other phrases captured include: "Later", "In response", "For this", and "Ultimately".

Verbal tense- and aspect-based (Reichenbach) ordering Following Chambers et al. (2014), we use deterministic rules to establish precedence between events that have certain verbal tense and aspect. These rules are derived from linguistic analysis of tense and aspect by (Reichenbach, 1947; Derczynski and Gaizauskas, 2013). Example (5) illustrates a case in which we can accurately infer order just from this information. Because *has been phosphorylated* has past tense and perfective aspect, this model concludes that it precedes *share* (present tense, simple aspect) and thus the binding of histone H2A.

(5) These [PTIP] proteins also share the ability to bind histone H2A (or H2AX in mammals) that has been phosphorylated. . . .

The logic determining which tense-aspect combinations receive which precedence relations is identical to CAEVO, which is possible because it is open source[8]. However, CAEVO operates over annotations that include gold tense and aspect values, whereas this model additionally detects tense and aspect using Odin rules before applying this logic.

3.2 Feature-based Models

Most instances of causal precedence cannot be captured with deterministic rules, because they lack explicit words, phrases, or syntactic structures that unambiguously mark the relation. Using a combination of the surface, syntactic, and taxonomic features outlined in Table 3, we trained a set of statistical classifiers to detect causal precedence relations between pairs of events in our corpus. For training and testing purposes, we treated any instance not labeled as either "E1 precedes E2" or "E2 precedes E1" as a negative example. We examined the following statistical models: a linear kernel SVM (Chang and Lin, 2011), logistic regression (Fan et al., 2008), and random forest[9] (Surdeanu et al., 2014). For the SVM and logistic regression (LR) models, we also compared the effects of L1 and L2 regularization.

3.3 Latent Representation Models

Due to the complexity of the task and variety of causal precedence instances encountered during the annotation process, it is unclear whether a linear combination of engineered features is sufficient for broad coverage classification. For this reason, we introduce a latent feature representation model using an LSTM (Hochreiter and Schmidhuber, 1997; Bergstra et al., 2010; Chollet, 2015) to capture underlying semantic features by incorporating long-distance contextual information and selectively persisting memory of previous event pairs to aid in classification.

The basic architecture is shown in Figure 5. The input to this model is the provenance of the relation, i.e., the whole text containing the two events and the text in between. Formally, this is represented as a concatenated sequence of 200 dimensional vectors where each vector in the sequence corresponds to a token in the minimal sentential span encompassing the event pair being classified. Intuitively, this LSTM "reads" the text from left to right and outputs a classification label from the set of three when done. We consider two variations of this model: the basic model (LSTM) with the vector weights for each token uninitialized and a second form (LSTM+P) where the vectors are initialized using pre-training. In the pre-training configuration, the vector weights are initialized using word embeddings generated by a word2vec (Mikolov et al., 2013; Řehůřek and Sojka, 2010) model trained on the full text of over 1 million biomedical papers taken from the Open Access subset of PubMed. Because the corpus is only 1000 annotations, it was thought that pre-training could improve prediction of causal precedence and guide the model with distributional semantic representations specific to this domain.

Building on this simple blueprint, we designed a three-pronged "pitchfork" (FLSTM) where the span of E1, the span of E2, and the minimal sentential span encompassing E1 and E2 each serve as a separate input, allowing the model to explicitly address each of them as well as discover how these three inputs relate to one another. This architecture is shown in Figure 6. Each input feeds into its own LSTM and corresponding dropout layer before the three forks are merged via a concatenation of tensors. Like the basic model, one version of the "pitchfork" is trained with vector weights initialized using the pre-trained word embeddings (FLSTM+P).

4 Results

We summarize the performance of all these models on the dataset previously introduced in Table 4. We report results using micro precision, recall, and F1 scores for each model. With fewer than 200 instances of causal precedence occurring in 1000 annotations, training and testing for both the feature-based classifiers and latent feature models was performed using stratified 10-fold cross validation. For the latent feature models, training was parameterized using a maximum of 100

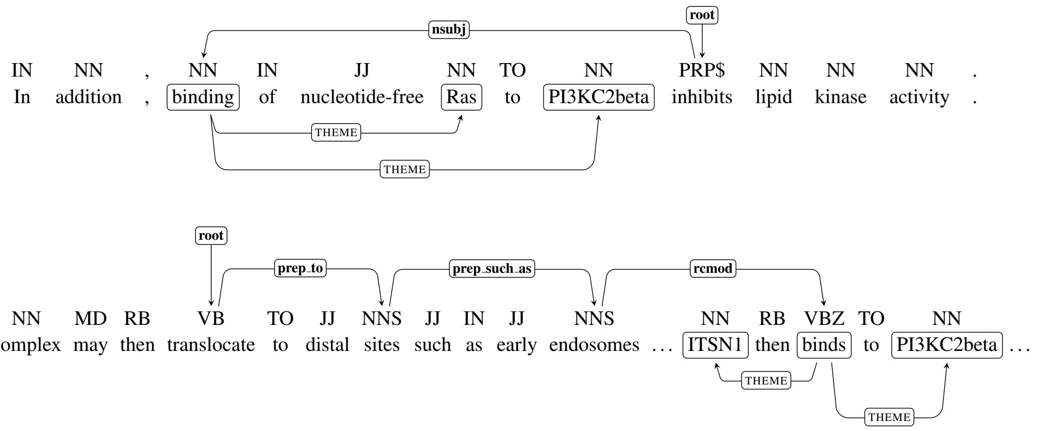
[8] `https://github.com/nchambers/caevo`

[9] Abbreviated as RF

	Feature	*Description*
Event	Event labels	The taxonomic labels Reach assigned to the event (e.g. *phosphorylation* → Phosphorylation, AdditiveEvent, . . .).
	Event trigger	The predicate signaling an event mention (ex. "phosphorylated", "phosphorylation").
	Event trigger + label	A concatenation of the event's trigger with the event's label.
	token *n*-grams with entity replacement	*n*-grams of the tokens in the mention span, where each entity is replaced with the entity label (ex. "the ABC protein" → "the PROTEIN"). If an entity is shared between pairs of events, replace it with the label SHARED.
	token *n*-grams with role replacement	*n*-grams of the tokens in the mention span, where each argument is replaced with the argument role (ex. "A inhibits the phosphorylation of B" → "CONTROLLER inhibits the CONTROLLED")
	Syntactic path from trigger to args	Variations of the syntactic dependency path from an event's trigger to each of its arguments (unlexicalized path, path + lemmas, trigger → argument role, trigger → argument label, etc.).
Event-Event (surface)	Interceding tokens (*n*-grams)	*n*-grams (1-3) of the tokens between E1 and E2.
Event-Event (syntax)	Cross-sentence syntactic paths	A concatenation of the syntactic path from the sentential ROOT to an event's trigger (see the example in Figure 4).
	Trigger-to-trigger syntactic paths (within sentence)	the syntactic path from the trigger of E1 to the trigger of E2
	Shortest syntactic paths	The shortest syntactic path between E1 and E2 (restricted to intra-sentence cases).
	Syntactic distance	The length of each syntactic path (restricted to intra-sentence cases).
Coreference	*Event* features for anaphors	Whether or not an event mention is resolved through coreference. For cases of coreference, generate the *Event* features prefixed with "coref-anaphor" for the text labeled "E1-anaphor" in the following example: (6) [A binds with B]$_{\text{E1-antecedent}}$ (7) [This interaction]$_{\text{E1-anaphor}}$ precedes the [phosphorylation of C]$_{\text{E2}}$
	Resolved arguments	Which arguments, if any, were resolved through coreference. For example: [The mutant$_{\text{THEME}}$ binds with B$_{\text{THEME}}$]$_{\text{E1}}$ → `THEME:resolved`

Table 3: An overview of the primary features used in the feature-based classifier, grouped into four classes: *Event* – features extracted from the two participating events, in isolation; *Event-Event (surface)* – features that model the lexical context between the two events; *Event-Event (syntax)* – features that model the syntactic context between the two events; and *Coreference* – features that capture coreference resolution information that impact the participating events.

In addition, **binding of nucleotide-free Ras to PI3KC2β** inhibits its lipid kinase activity. The PI3KC2β and Ras complex may then translocate to distal sites such as early endosomes (EE) where **ITSN1 then binds to PI3KC2β** leading to the release of nucleotide-free Ras and activation of the lipid kinase activity of PI3KC2β.


root
nsubj
IN NN , NN IN JJ NN TO NN PRP$ NN NN NN .
In addition , binding of nucleotide-free Ras to PI3KC2beta inhibits lipid kinase activity .
THEME
THEME
root
prep_to
prep_such_as
rcmod
NN MD RB VB TO JJ NNS JJ IN JJ NNS NN RB VBZ TO NN
. . . complex may then translocate to distal sites such as early endosomes . . . ITSN1 then binds to PI3KC2beta . . .
THEME
THEME

ROOT >NSUBJ + ROOT >PREP_TO >PREP_SUCH_AS >RCMOD

Figure 4: Generation procedure for the cross-sentence syntactic path feature. For each event in a pair, we find the shortest syntactic path originating from the sentential root node leading to a token in the event's trigger. The two syntactic paths are then joined using the + symbol to form a single feature.

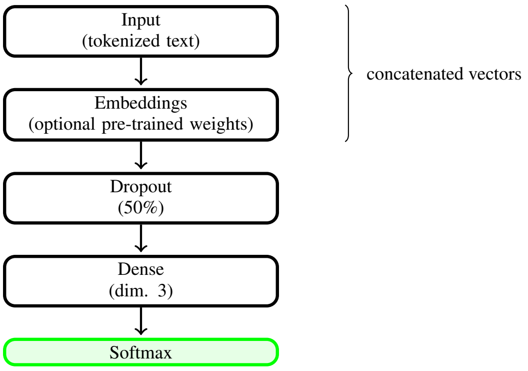


Figure 5: Architecture for the basic latent feature model using the minimal sentential span encompassing events 1 and 2 as input.

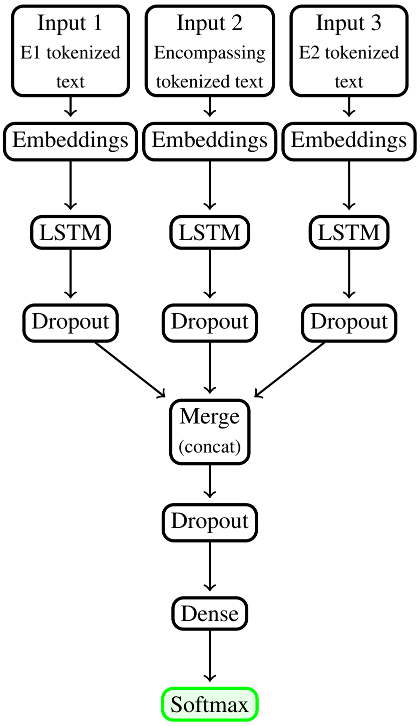


Figure 6: Modified architecture for a latent feature model with three-pronged input: the text of event 1 (left), the minimal sentential span encompassing events 1 and 2 (middle), and the text of event 2 (right).

epochs with support for early stopping through monitoring of validation loss[10]. Weight updates were made on batches of 32 examples and all folds completed in fewer than 50 epochs.

The table also includes a sieve-based ensemble system, which performs significantly better than the best-performing single model. In this architecture, the sieves are applied in descending order of precision, so that the positive predictions of the higher precision sieves will always be preferred to contradictory predictions made by subsequent, lower-precision sieves. Figure 7 illustrates that as sieves are added, the F1 score remains fairly constant, while recall increases at the cost of precision.

Model	*p*	*r*	*f1*
Intra-sentence	0.5	0.01	0.01
Inter-sentence	0.5	0.01	0.01
Reichenbach	0	0	0
LR+L1	0.58	0.32	0.41
LR+L2	0.65	0.26	0.37
SVM+L1	0.54	0.35	**0.43**
SVM+L2	0.54	0.29	0.38
RF	0.62	0.25	0.36
LSTM	0.40	0.25	0.31
LSTM+P	0.39	0.20	0.26
FLSTM	0.43	0.15	0.22
FLSTM+P	0.38	0.22	0.28
Combined	0.38	0.58	**0.46***

Table 4: Results of all proposed causal models, using stratified 10-fold cross-validation. The combined system is a sieve-based architecture that applies the models in decreasing order of their precision. The combined system significantly outperforms the best single model, SVM with L1 regularization, according to a bootstrap resampling test (p = 0.022).

Despite some obvious patterns noted in Table 1, the deterministic models perform the worst due in large part to their rarity in the corpus. An analysis of this result is given in Section 5. Overall, our top-performing model was the linear kernel SVM with L1 regularization. In all cases, the feature-based classifiers outperform the latent feature representations, suggesting that in cases such as this where little data is available, feature-based classifiers capitalizing on high-level linguistic features are able to better generalize and thus outperform latent feature models. However, as our discussion in Section 5.1 will show, our combined model demonstrates that the latent and feature-based models are largely complementary.

[10] The validation set used for each fold came from a different class-balanced fold.

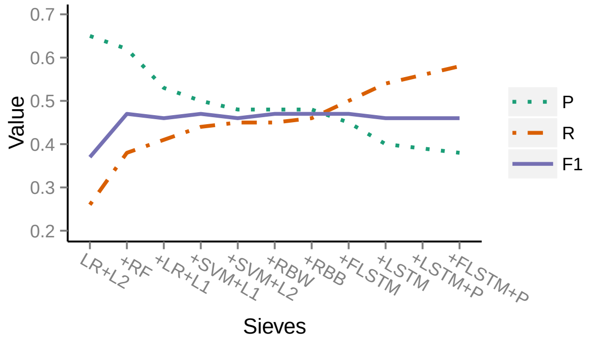


Figure 7: The performance of the sieve-based combined model varies with each model added.

5 Discussion

Overall, results are promising, particularly in light of the conscious choice to omit (causal) regulation reactions from this task, as they are already captured by the Reach reading system.

However, the deterministic models created so far have extremely low recall, such that it is difficult even to determine their precision. An analysis of the Reichenbach model reveals one source of this low coverage. In short, although writers *could* describe causal mechanisms using temporal indicators such as tense and aspect, temporal description is rare enough in this domain not to be represented in our randomly sampled database. Table 5 illustrates the lack of overlap with informative tense-aspect combinations; a single tense is used per passage, and no perfective aspect is used.

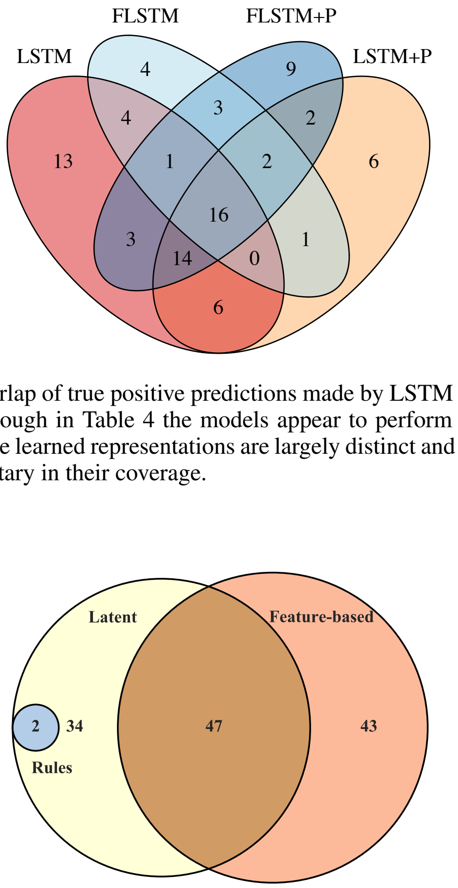
E1↓, E2→		*past*		*pres.*		*fut.*	
		simple	*perf.*	*simple*	*perf.*	*simple*	*perf.*
past	*simple*	69	0	38	0	0	0
	perf.	0	0	0	0	0	0
pres.	*simple*	49	0	134	0	1	0
	perf.	0	0	0	0	0	0
fut.	*simple*	0	0	0	0	0	0
	perf.	0	0	0	0	0	0

Table 5: Event tense and aspect for events containing verbs in the present study. Highlighted cells are tense-aspect combinations that are informative for establishing temporal precedence, following Chambers et al. (2014). All but one event pair fall outside of these informative combinations, and that exceptional pair was a false positive case.

Similarly, the time expressions required by the deterministic intra- and inter-sentence precedence rules are rare enough to make them ineffective on this sample.

5.1 Model overlap

As Chambers et al. (2014), Mirza (2016), and many other algorithms have shown, models can be applied sequentially in "sieves" to produce higher-quality output. Ideally, each model in a sieve-based system will capture different portions of the data through a mixture of approaches, distinguishing this method from more naive ensembles in which the contributions of a lone component would be washed out. Figure 8 details this observation by showing the coverage difference between the models described here.



(a) Overlap of true positive predictions made by LSTM models. Though in Table 4 the models appear to perform similarly, the learned representations are largely distinct and complementary in their coverage.

(b) Similarly, the overlap between the feature-based models and the latent models was low overall.

Figure 8: The overlap of true positives among the investigated models was low.

5.2 Error analysis

We performed an analysis of the false positives shared by all feature-based classifiers, in addition to the false negatives shared by all models. Here we limit our discussion to only the most prominent characteristic shared by the majority of false positives.

Discourse information More than half of the false positives share contrastive discourse features, suggesting that a model of discourse could improve classifier discrimination. Example (8) demonstrates such a contrastive structure, which *whereas* introduces a clause (and event) that is contrasted and therefore both temporally and causally distinct from the following clause (and event). The existence of regular cues like *whereas* indicates that a feature to explicitly model these structures is possible.

(8) Whereas [PRAS40 inhibits the mTORC1 activity via raptor]$_{\mathrm{E1}}$, DEPTOR was identified to interact directly with mTOR in both [mTORC1 and mTORC2 complexes]$_{\mathrm{E2}}$

6 Related Work

Though focused on temporal ordering, Chambers et al. (2014) adopt a sieve-based approach, with high-precision deterministic sieves preceding and constraining lower-precision, higher-recall machine learning sieves. As with our system, the deterministic sieves were linguistically motivated, and had the additional advantage of operating over time expressions (*during*, *Friday*, etc.) as well as events, the former of which are typically lacking in the biomedical domain.

Mirza (2016) implemented a hybrid sieve-based approach for causal relation detection between events that includes a set of causal verb rules and corresponding syntactic dependencies and a feature-based classifier. However, both of these works focus on open-domain texts. To our knowledge, we are the first to investigate causal precedence in the biomedical domain.

7 Conclusion

These are the first experiments regarding automatic annotation of causal precedence in the biomedical domain. Although the dearth of temporal expressions and other regular linguistic cues make the task especially difficult in this domain, the initial results are promising, and demonstrate that a sieve-based system of the models tested here improves performance over the top-performing individual component. Both the annotation corpus and the models described here represent large steps toward linking automatic reading to a larger, more informative biological mechanism.

Acknowledgments

This work was funded by the Defense Advanced Research Projects Agency (DARPA) Big Mechanism program under ARO contract W911NF-14-1-0395.

References

Dane Bell, Gus Hahn-Powell, Marco A. Valenzuela-Escárcega, and Mihai Surdeanu. 2016. An investigation of coreference phenomena in the biomedical domain. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*. LREC 2016. Paper available at `http://arxiv.org/abs/1603.03758`.

James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. 2010. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June. Oral Presentation.

Steven Bethard, Leon Derczynski, Guergana Savova, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. Semeval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical tempeval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, California, June. Association for Computational Linguistics*.

Ethan G Cerami, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Özgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. 2011. Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, 39(suppl 1):D685–D690.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Franois Chollet. 2015. keras. `https://github.com/fchollet/keras`.

Jacob Cohen. 1960. A coefficient of agreement for nominal scale. *Educational and Psychosocial Measurement*, 20:37–46.

Paul R. Cohen. 2015. DARPA's Big Mechanism program. *Physical Biology*, 12(4):045008.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. Stanford typed dependencies manual. Technical report, Stanford University.

Leon Derczynski and Robert Gaizauskas. 2013. Empirical validation of Reichenbach's tense framework. In *Proceedings of the 10th International Conference on Computational Semantics*, pages 71–82.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9. Association for Computational Linguistics.

Jin-Dong Kim, Ngan Nguyen, Yue Wang, Junichi Tsujii, Toshihisa Takagi, and Akinori Yonezawa. 2012. The genia event and protein coreference tasks of the bionlp shared task 2011. *BMC bioinformatics*, 13(11):1.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Claudiu Mihăilă, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. Biocause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14(1):1–18.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Paramita Mirza. 2016. Extracting temporal and causal relations between events.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Hans Reichenbach. 1947. *Elements of Symbolic Logic*. Macmillan, London.

Mihai Surdeanu, Marco Valenzuela-Escárcega, Gus Hahn-Powell, Peter Jansen, Daniel Fried, Dane Bell, and Tom Hicks. 2014. processors. `https://github.com/clulab/processors`.

Marco Valenzuela-Escárcega, Gus Hahn-Powell, Dane Bell, Tom Hicks, Enrique Noriega, and Mihai Surdeanu. 2015a. Reach. `https://github.com/clulab/reach`.

Marco A. Valenzuela-Escárcega, Gus Hahn-Powell, and Mihai Surdeanu. 2015b. Description of the odin event extraction framework and rule language.

Marco A. Valenzuela-Escárcega, Gustave Hahn-Powell, Thomas Hicks, and Mihai Surdeanu. 2015c. A domain-independent rule-based framework for event extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing: Software Demonstrations (ACL-IJCNLP)*, pages 127–132. ACL-IJCNLP 2015.