Proceedings of the Ninth Annual Conference on Advances in Cognitive Systems Submitted 9/2021; published 11/2021

From Unstructured Text to Causal Knowledge Graphs: A Transformer-Based Approach

Scott Friedman FRIEDMAN@SIFT.NET
Ian Magnusson IMAGNUSSON@SIFT.NET
Vasanth Sarathy VSARATHY@SIFT.NET
Sonja Schmer-Galunder SGALUNDER@SIFT.NET
SIFT, 319 N 1st Ave., Minneapolis, MN 55401 USA

Abstract

Qualitative causal relationships compactly express the direction, dependency, temporal constraints, and monotonicity constraints of discrete or continuous interactions in the world. In everyday or academic language, we may express interactions between quantities (e.g., sleep decreases stress), between discrete events or entities (e.g., a protein inhibits another protein's transcription), or between intentional or functional factors (e.g., hospital patients pray to relieve their pain). Extracting and representing these diverse causal relations are critical for cognitive systems that operate in domains spanning from scientific discovery to social science. This paper presents a transformer-based NLP architecture that jointly extracts knowledge graphs including (1) variables or factors described in language, (2) qualitative causal relationships over these variables, (3) qualifiers and magnitudes that constrain these causal relationships, and (4) word senses to localize each extracted node within a large ontology. We do not claim that our transformer-based architecture is itself a cognitive system; however, we provide evidence of its accurate knowledge graph extraction in real-world domains and the practicality of its resulting knowledge graphs for cognitive systems that perform graph-based reasoning. We demonstrate this approach and include promising results in two use cases, processing textual inputs from academic publications, news articles, and social media.

1. Introduction

People express causal relationships in everyday language and scientific texts to capture the relationship between quantities or entities or events, compactly communicating how one event or purpose or quantity might affect another. These causal relations are not complete mechanisms in themselves, but we use them frequently in everyday language and formal instruction to express causality, allowing us to avoid unnecessary detail or to hedge when details are uncertain.

Identifying these causal relationships from natural language—and also properly identifying the factors that they relate—remains a challenge for cognitive systems. This difficulty is due in part to the expressiveness of our language, e.g., the multitude of ways we may describe how an experimental group scored higher on an outcome than a control group, and also due to the complexity of the systems we describe.

This paper describes an approach to automatically extracting (1) entities that are the subject of causal relationships, (2) causal relationships describing mechanisms, intentions, monotonicity, and temporal priority, (3) multi-label attributes to further characterize the causal structure, and (4) ontologically-grounded word senses for applicable nodes in the causal graph. Our primary claim is that context-sensitive language models can detect and characterize the qualitative causal structure of everyday and scientific language in a representation that is usable by cognitive systems. As evidence, we present our SpEAR transformer-based NLP model based on BERT (Devlin et al., 2019) and SpERT (Eberts & Ulges, 2020) that extracts causal structure from text as knowledge graphs, and we present promising initial results on (1) characterizing scientific claims and (2) representing and traversing descriptive mental models from ethnographic texts.

The present work aims to infer causal, functional, and intentional relational structure, so its output knowledge representations are relevant to cognitive systems; however, the NLP methodology that performs the inference is not intended to model human cognition. The nodes within the causal, semantic graphs produced by SpEAR link to the WordNet word sense hierarchy (Fellbaum, 2010) to facilitate subsequent reasoning. Unlike rule-based parsers that use ontological constraints during the parsing process, our NLP architecture infers ontological labels (i.e., WordNet word senses) as a context-sensitive post-process. We demonstrate that the knowledge representations inferred by our system allows traversal across concepts to characterize meaningful causal influences.

We continue with a review of related work in qualitative causal representations (Section 2.1) and transformer-based NLP (Section 2.2). We then describe our approach (Section 3) and preliminary results in two domains (Section 4). We conclude with a discussion of future work in this area.

2. Background and Related Work

We review related work in representing causal relations, which informs the present approach. We then review previous work in transformer-based NLP, including the SpERT system (Eberts & Ulges, 2020) which is a subsystem of our architecture.

2.1 Qualitative Causal Relations

The knowledge representations described in this paper are motivated by previous work in qualitative reasoning and simulation (Forbus, 2019). For example, *qualitative proportionalities* describe how one quantity impacts another, in a directional, monotonic fashion. In this work, we designate $\langle a, \textbf{q+}, b\rangle$ (and respectively, $\langle a, \textbf{q–}, b\rangle$) as qualitative proportionalities from a to b, such that increasing a would increase (and respectively, decrease) b. This is motivated by quantity-to-quantity $\alpha_{Q+/-}$ relations (Forbus, 1984) and $M^{+/-}$ relations in qualitative simulation (Kuipers, 1986). Our semantics are less constrained than either of these, due to tendencies in language to express an increase from an event to a quantity (e.g., “smoking a cigarette will increase your risk of cancer”) or from entities to activities (e.g., “the prime increased participants’ retrieval of the cue”), and so on.

Previous work in philosophy (Dennett, 1989) and cognitive psychology (Lombrozo & Carey, 2006) has acknowledged intentional (i.e., psychological, goal-based) and teleological (i.e., func-

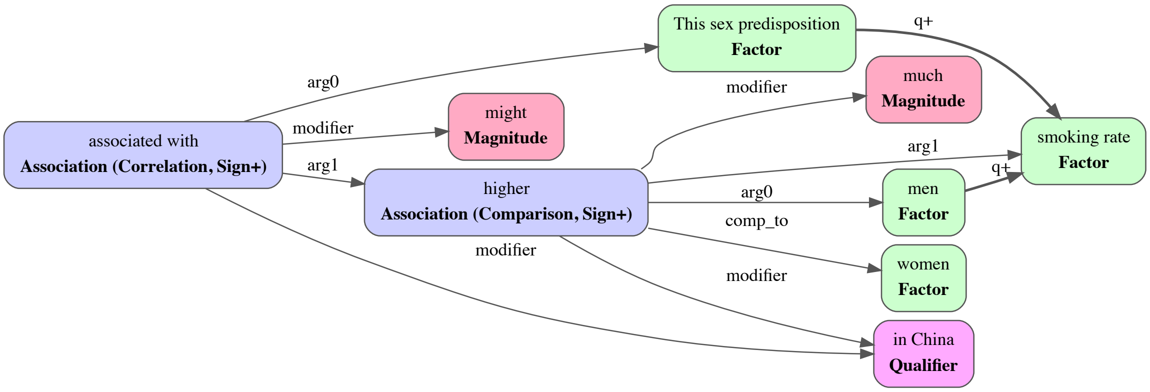

Figure 1: SpEAR knowledge graph output for the text “This sex predisposition might be associated with the much higher smoking rate in men than in women in China.” Includes a correlation, a comparison with a qualitative increase, magnitudes, and a location qualifier.

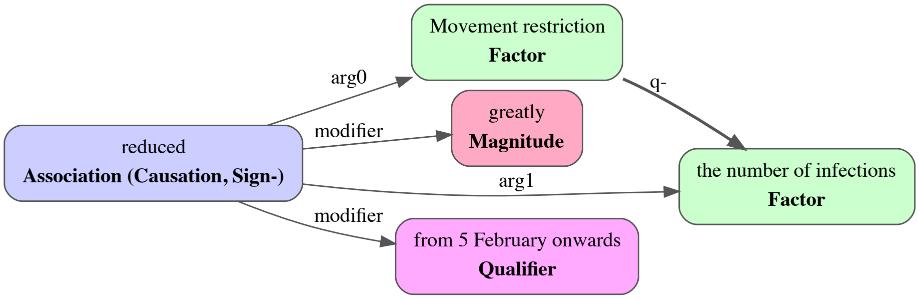

Figure 2: SpEAR knowledge graph output for “Movement restriction greatly reduced the number of infections from 5 February onwards.” Includes a causal association, a qualitative decrease, a magnitude, and a temporal qualifier.

tional, design-based) relationships as types of causal relations. Previous work has represented these as lexical qualia or affordances (Pustejovsky, 1991). In this work, we represent purposeful, intentional actions as a qualitative relationship $\langle a, \textbf{intent+}, b\rangle$, such that the actor of action a may have intended the purpose or goal b, e.g., “they prayed for a safe pregnancy.” We represent teleological (i.e., functional or design-based) causal relations as $\langle a, \textbf{function+}, b\rangle$ to indicate that the action or artifact a is designed or otherwise has a function to achieve b, e.g., “the artifacts provide protection for pregnant women.”

2.2 Causal and Transformer-Based NLP

Transformer-based methods for NLP utilize neural networks to encode a sequence of textual tokens (i.e., words or sub-words) into large vector-based representations for each token, sensitive to the context of the surrounding tokens (Devlin et al., 2019). This is widely regarded as a state-of-theart methodology for NLP, and has been used to process text to extract knowledge graphs, e.g., of people and relations (Eberts & Ulges, 2020). The architecture we present in this paper has been applied to the SciClaim dataset of scientific claims (Magnusson & Friedman, 2021) and a social media corpus centered on moral attributions and hate speech (Friedman et al., 2021). Many existing transformer models—similar to the architecture presented in this paper—require hundreds (sometimes thousands) of labeled training examples to reach high proficiency.

Existing symbolic semantic parsers extract scientific claims and assertions from text with explicit relational knowledge representations (Allen et al., 2015). Many of these rely on rule-based engines with hand tuning, which provides more customization and interpretability, at the expense of using NLP experts to maintain and adapt to new domains. By contrast, our approach extracts causal knowledge graphs using advances in transformer-based models such as SpERT (Eberts & Ulges, 2020) to learn graph-based representations from examples alone. The resulting knowledge graphs are ontologically-grounded and support graph-based reasoning, as we demonstrate below; however, these are not as expressive as some modern symbolic parsers.

Other NLP approaches use machine learning to extract features from scientific texts, e.g., to identify factors and directions of influence in scientific claims (Mueller & Abdullaev, 2019); however these approaches do not explicitly infer relations between elements in a causal graph or the ontological groundings of the terms, as in our approach.

3. Approach

We describe our graph schema for representing the entities, attributes, and qualitative relationships extracted from text. We discuss the general problem definition and then we explain the specific graph schemas in two domains: (1) scientific claims and (2) ethnographic mental models.

3.1 Knowledge Graphs

The SpEAR knowledge graph format includes the following three types of elements: entities, attributes, and relations. We describe each of these before defining the problem and describing the architecture.

Entities. Entities are labeled spans within a textual example. These are the nodes in the knowledge graph. The same exact span cannot correspond to more than one entity type, but two entity spans can overlap. Entities comprise the nodes of Figures 1-3 upon which attributes and relations are asserted. Unlike most ontologically-grounded symbolic parsers (e.g., Das et al., 2010; Allen et al., 2015), these entity nodes are not ontologically grounded in a class hierarchy; rather, these entity nodes are associated with a token sequence (e.g., “smoking rate” in Figure 1) and a corresponding entity class (e.g., **Factor**). These entities also have high-dimensional vectors from the transformer model, which approximates the distributed semantics. Our architecture also associates entities with applicable WordNet senses, as we describe below in Section 3.4.7.

Attributes. Attributes are Boolean labels, and each entity (i.e., graph node) may have zero or more associated attributes. Attribute inference is therefore a multi-label classification problem. The previous SpERT transformer model was not capable of expressing these; this is a novel contribution of SpEAR, as described in Section 3.4. In Figures 1-3, attributes are rendered as parenthetical labels inside the nodes, e.g., **Correlation** and **Sign+** in the Figure 1 nodes for "associated with" and "higher," respectively. The multi-label nature allows the Figure 1 "higher" node to be categorized simultaneously as **Sign+** and **Comparison**.

Relations. Relations are directed edges between labeled entities, representing semantic relationships. These are critical for expressing what-goes-with-what over the set of entities. For example in the sentence in Figure 1, the relations (i.e., edges) indicate that the "higher" association asserts the antecedent (**arg0**) "men" against (**comp_to**) "women" for the consequent (**arg1**) "smoking rate." In Figures 1-3, the **modifier** relations link nodes to others that semantically modify them. Without all of these labeled relations, the semantic structure of these scientific claims would be ambiguous.

3.2 Problem Definition

We define the multi-attribute knowledge graph extraction task as follows: for a text passage $\mathcal{S}$ of n tokens $s_1, ..., s_n$, and a schema of entity types $\mathcal{T}_e$, attribute types $\mathcal{T}_a$, and relation types $\mathcal{T}_r$, predict:

1. The set of entities $\langle s_j, s_k, t \in \mathcal{T}_e \rangle \in \mathcal{E}$ ranging from tokens s_j to s_k, where $0 \leq j \leq k \leq n$,
2. The set of relations over entities $\langle e_{head} \in \mathcal{E}, e_{tail} \in \mathcal{E}, t \in \mathcal{T}_r \rangle \in \mathcal{R}$ where $e_{head} \neq e_{tail}$,
3. The set of attributes over entities $\langle e \in \mathcal{E}, t \in \mathcal{T}_a \rangle \in \mathcal{A}$.

This defines a directed multi-graph without self-cycles, where each node has zero to $|\mathcal{T}_a|$ attributes. SpEAR does not presently populate attributes on relations.

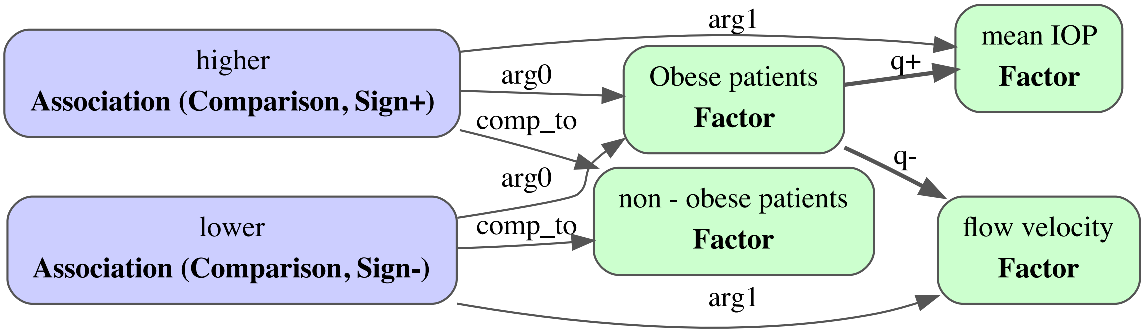

Figure 3: SpEAR knowledge graph output for "Obese patients have a higher mean IOP and lower flow velocity than non-obese patients." The two qualitative comparisons "higher" and "lower" support qualitative **Sign+** and **Sign-** attributes, and **q+** and **q-** relations, respectively.

3.3 Knowledge Graph Schemas

We briefly describe a subset of the graph schemas for our two use-cases: (1) the SciClaim dataset of scientific claims and (2) ethnographic mental models. These two schemas share some qualitative causal representations but vary in other domain-specific descriptions. In follow-on work, these schemas may be integrated into a single schema.

Scientific Claims. The SciClaim scientific claim schema is designed to capture associations between factors (e.g., causation, comparison, prediction, proportionality), monotonicity constraints across factors, epistemic status, and high-level qualifiers. This model is used for qualitative reasoning to help characterize the replicability and reproducibility of scientific claims (Alipourfard et al., 2021; Gelman et al., 2021). We describe the entities, attributes, and relations of the schema, referencing the graphed examples rendered by our system in Figures 1, 2, and 3.

This schema includes six entity types: **Factors** are variables that are tested or asserted within a claim (e.g., “smoking rate” in Figure 1); **Associations** are explicit phrases associating one or more factors in a causal, comparative, predictive, or proportional assertion (e.g., “associated with” and “reduced” in Figures 1 and 2, respectively); **Magnitudes** are modifiers of an association indicating its likelihood, strength, or direction (e.g., “might” and “much” in Figure 1); **Evidence** is an explicit mention of a study, theory, or methodology supporting an association; **Epistemics** express the belief status of an association, often indicating whether something is hypothesized, assumed, or observed; **Qualifiers** constrain the applicability or scope of an assertion (e.g., “in China” in Figure 1 and “from 5 February onwards” in Figure 2).

This schema includes the following attributes, all of which apply solely to the *Association* entities: **Causation** expresses cause-and-effect over its constituent factors (e.g., “reduced” span in Figure 2); **Comparison** expresses an association with a frame of reference, as in the “higher” statement of Figure 1 and the “higher” and “lower” statements of Figure 3; **Sign+** expresses high or increased factor value; **Sign-** expresses low or decreased factor value; **Indicates** expresses a predictive relationship; and **Test** indicates a statistical test employed to test a hypothesis.

We encode six relations: **arg0** relates an association to its cause, antecedent, subject, or independent variable; **arg1** relates an association to its result or dependent variable; **comp_to** is a frame of reference in a comparative association; **modifier** relates entities to descriptive elements; **q+** and **q-** indicate positive and negative qualitative proportionality, respectively, where increasing the head factor increases or decreases (the amount or likelihood of) the tail factor, respectively.

Ethnographic Mental Models. In our preliminary ethnographic mental modeling domain, we utilize a slightly different schema to capture intentional and functional causality in addition to culturally-specific attributes such as gender and spirituality. We use Figure 4 and Figure 5 to illustrate the ethnographic causal graph schema.

This schema includes attributes for spiritual or cultural **Tradition** (e.g., “prayed” in Figure 4), **Event** (e.g., “gave” and “drink” in Figure 5), **Influence** for causally-potent elements (e.g., “prevent” in Figure 4), and others.

We include additional relations **agent**, **object**, **recipient**, **consequent**, and others as semantic role relations of events and assertions. These relations (rendered in narrow lines in Figure 4 and Figure 5) comprise a description logic of their head nodes, such that the head node would not have the same semantics without the its reachable subgraph along these edges.

The bold-rendered edges are causal edge, including qualitative monotonicity **q+** and **q-**, temporal precedence **t+** relations to indicate one event preceding another, and intentional **intent+** and functional **function+** relations to indicate the goal (i.e., intention or function, respectively) of an action or artifact. For instance, the graph in Figure 4 shows an **intent+** from “prayed” to “prevent” and then a **q-** to “complications”, ultimately indicating that prayer has a goal of minimizing compli-

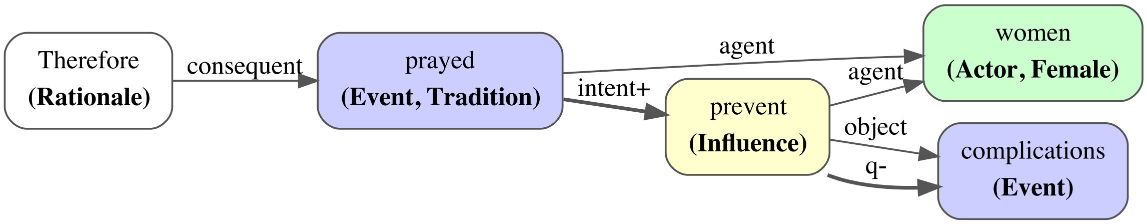

Figure 4: SpEAR knowledge graph for "Therefore, the women prayed to prevent any complications," including **intent+** and **q-** relations.

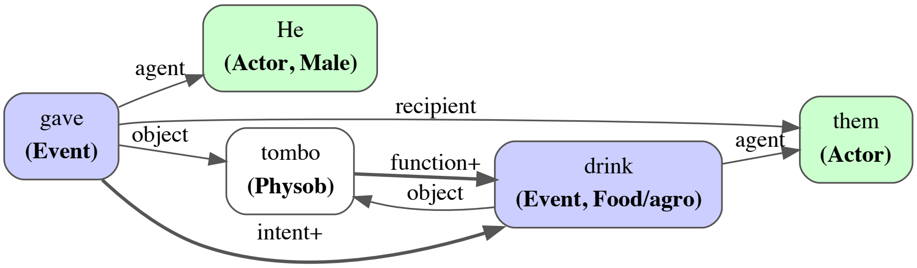

Figure 5: SpEAR knowledge graph for "He also gave them tombo to drink.", including **intent+** and **function+** relations.

cations. Furthermore, the graph in Figure 5 illustrates an **intent+** relation from "gave" to "drink," indicating the giving is intended to support the drinking. Figure 5 also includes a **function+** relation, indicating that the "tombo" is designed or cultivated for drinking.

The relatively simple statement in Figure 4 originates from an ethnographic article (Aziato et al., 2016) that includes interview snippets, and the sentence in Figure 5 is from a collection of international folktales.[1] Despite their simplicity, the SpEAR knowledge graphs illustrate rich multi-step causality: Figure 4 indicates that prayer has the purpose of reducing the incidence (or severity of) complications, and Figure 5 plots a simple narrative structure indicating an agent's intention to affect the actions of other agents, as well as the function of a novel entity.

3.4 Model Architecture

Our SpEAR model architecture extends SpERT with an attribute classifier and attention-based span representation. The original architecture provides components (Figure 6 a–c) for joint entity and relation extraction on potentially-overlapping text spans. The parameters of the entity, attribute, and relation classifiers, as well as the parameters of the BERT language model (initialized with its pre-trained values) are all trained end-to-end on our dataset.

[1] https://www.worldoftales.com/

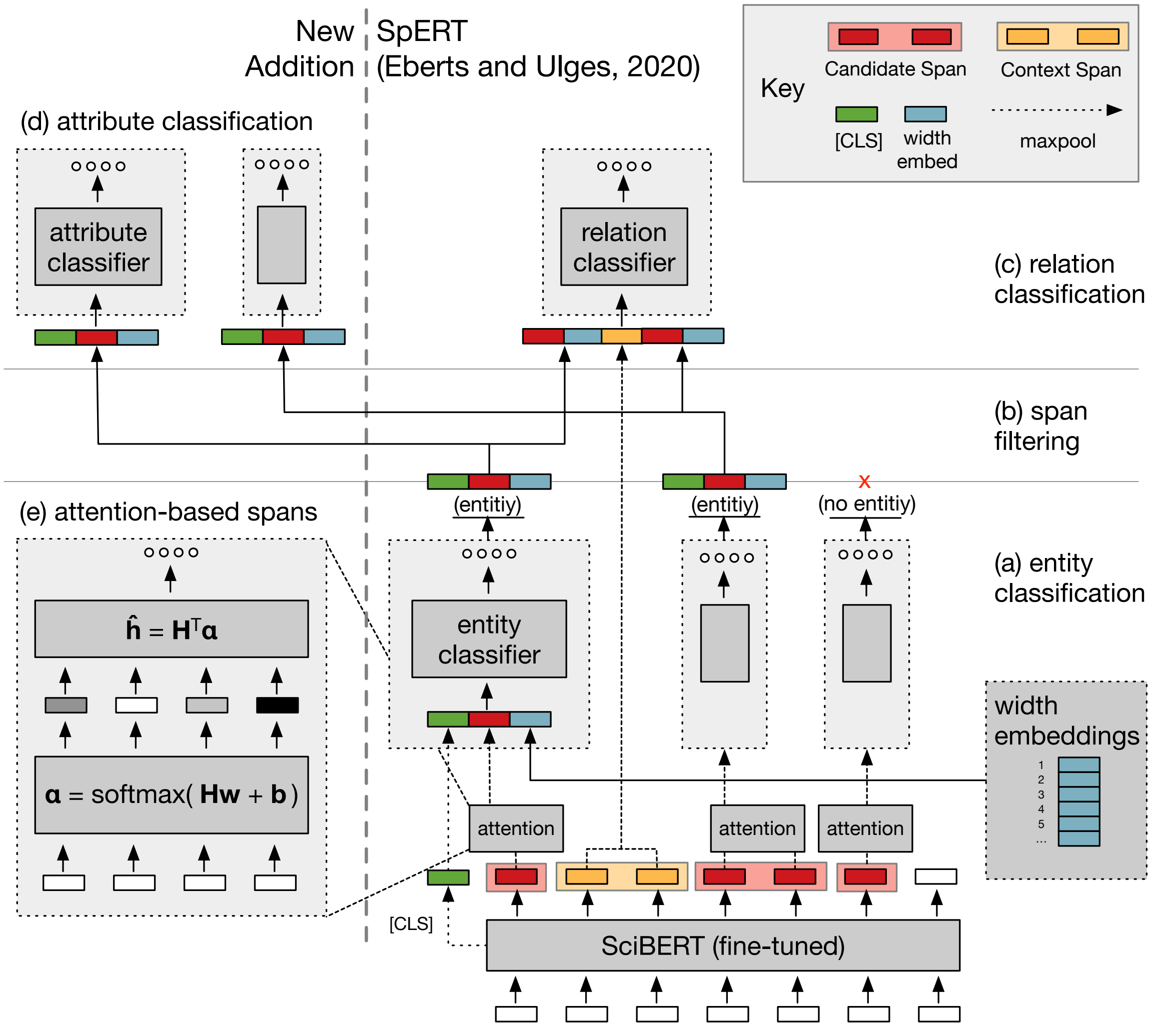

Figure 6: The SpEAR transformer-based model extends the SpERT components (a, b, and c) with attribute classification (d) that performs multi-label inference on identified entity spans and attention-based representations (e) of spans, inspired by Lee et al. (2017).

3.4.1 Computing Token Vectors

The tokens $s_1, ..., s_n$ of the text passage S are each embedded by a transformer such as BERT (Devlin et al., 2019) as a sequence $\mathbf{e}_1, ..., \mathbf{e}_n$ of high-dimensional vectors representing the token and its context. BERT also provides an additional "[CLS]" vector output, $\mathbf{e}_0$, designed to represent information from the complete text input. For all possible spans, $span_{j,k} = s_j, ..., s_k$, up to a given length, the word vectors associated with a span, $\mathbf{e}_j, ..., \mathbf{e}_k$, are combined into a final span vector, $\mathbf{e}(span_{j,k})$.

3.4.2 Computing Span Vectors

The original SpERT architecture uses *maxpooling* to compute each dimension of $\mathbf{e}(span_{j,k})$ as the maximum value across its constituent BERT token vectors for that dimension. Instead of using maxpool, SpEAR uses an attention-based span representation (Figure 6e) inspired by Lee et al. (2017) to compute span vectors. This produces *attention weight* scalars $\alpha_{i,t}$ for each BERT token vector $\mathbf{h}_t$ in a span i using learned parameters $\mathbf{w}$ and b:

$$\alpha_{i,t} = \frac{\exp(\mathbf{w} \cdot \mathbf{h}_t + b)}{\sum_{k=START(1)}^{END(i)} \exp(\mathbf{w} \cdot \mathbf{h}_k + b)} \tag{1}$$

These attention weights help compute the span representation $\hat{\mathbf{h}}_i$ with the following weighted sum:

$$\hat{\mathbf{h}}_i = \sum_{t=START(1)}^{END(i)} \alpha_{i,t}\mathbf{h}_t \tag{2}$$

3.4.3 Classifying Spans as Entities

The final attention-based span representation, $\mathbf{x}(span_{j,k})$ is made by concatenating together the attention representation $\mathbf{e}(span_{j,k})$ and $\mathbf{e}_0$ along with a width embedding, $\mathbf{w}_l$, that encodes the number of words, l, in $span_{j,k}$. Each valid span length l looks up a different vector of learned parameters, $\mathbf{w}_l$. SpEAR uses the concatenated $\mathbf{x}(span_{j,k})$ vector to classify spans into mutually-exclusive entity types (including a *null* type) using a linear classifier (Figure 6a). Only spans identified as entities move on to further analysis (Figure 6b).

3.4.4 Inferring Multi-Class Attributes on Entities

SpEAR uses its classified entities $\mathbf{x}^a$ as inputs to its attribute classifier (Figure 6d) with weights $\mathbf{W}^a$ and bias $\mathbf{b}^a$. A pointwise sigmoid σ yields separate confidence scores $\hat{\mathbf{y}}^a$ for each attribute in the graph schema:

$$\hat{\mathbf{y}}^a = \sigma(\mathbf{W}^a\mathbf{x}^a + \mathbf{b}^a) \tag{3}$$

We train the attribute classifier with a binary cross entropy loss $\mathcal{L}_a$ summed with the SpERT entity and relation losses, $\mathcal{L}_e$ and $\mathcal{L}_r$, for a joint loss:

$$\mathcal{L} = \mathcal{L}_e + \mathcal{L}_r + \mathcal{L}_a \tag{4}$$

SpEAR takes only identified entity spans as input to the attribute classifier, as this approach provided best performance and aligns with the finding by Eberts & Ulges (2020) that training on downstream tasks yields best accuracy with strong negative samples of ground truth entities (i.e., teacher forcing).

3.4.5 Inferring Labeled Relations between Entities

SpEAR uses all pairings of classified entities (Figure 6b) as inputs to its relational classifier (Figure 6c). SpEAR's relational classifier identical to SpERT's: a multi-label linear classifier that takes each pair of entities (i.e., a relation head and a relation tail) and concatenates their span representations, width representations, and also the maxpool of the token vectors between the two entities. The output of the relational classifier is zero or more labeled relations from the head entity to the tail entity.

The output of SpEAR's neural components comprises a *directed multigraph* (i.e., a directed graph that is allowed to have multiple edges between any two nodes) without self-loops. The multigraph may be disconnected, and may contain isolated nodes. Each node (i.e., labeled entity) in the multigraph may have zero or more Boolean attributes. Every entity, attribute, and relation in SpEAR's directed multigraphs includes a *confidence score* between 0 and 1.

3.4.6 Rectifying Results

SpEAR includes a novel *rectifier* component (not shown in Figure 6) that prunes entities, attributes, and relations that are inconsistent with the constraints of the graph schema. For example, relations might be constrained to originate or terminate at certain entity types, attributes may be constrained to certain entity types, and some attributes and relations may be mutually inconsistent.

When the rectifier detects a schema conflict, it uses SpEAR's confidence scores to remove lower-score elements to resolve it. This strictly removes graph elements, so it cannot improve SpEAR's recall score— and it may even reduce the recall score— but empirically, we find the rectifier increases precision proportionately and ultimately increases SpEAR's F1 measure in some domains.

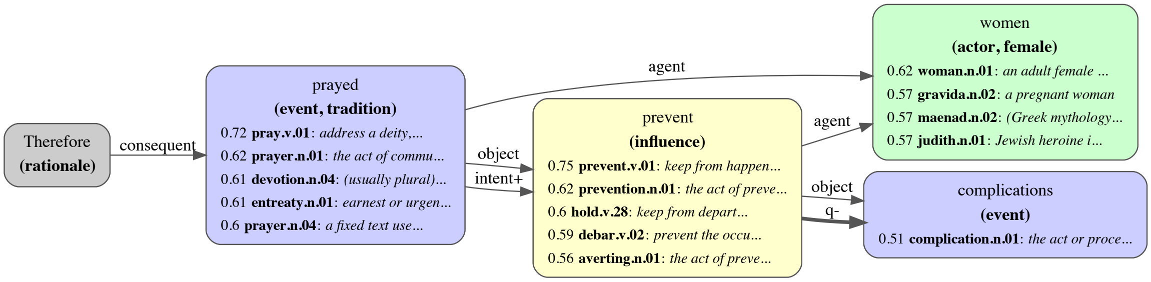
Therefore
(rationale)
consequent
prayed
(event, tradition)
0.72 pray.v.01: address a deity,...
0.62 prayer.n.01: the act of commu...
0.61 devotion.n.04: (usually plural)...
0.61 entreaty.n.01: earnest or urgen...
0.6 prayer.n.04: a fixed text use...
agent
object
intent+
prevent
(influence)
0.75 prevent.v.01: keep from happen...
0.62 prevention.n.01: the act of preve...
0.6 hold.v.28: keep from depart...
0.59 debar.v.02: prevent the occu...
0.56 averting.n.01: the act of preve...
agent
object
q-
women
(actor, female)
0.62 woman.n.01: an adult female ...
0.57 gravida.n.02: a pregnant woman
0.57 maenad.n.02: (Greek mythology...
0.57 judith.n.01: Jewish heroine i...
complications
(event)
0.51 complication.n.01: the act or proce...

Figure 7: SpEAR knowledge graph for the same sentence in Figure 4, also displaying WordNet word senses automatically inferred by the architecture. The listed word senses include a confidence score, the WordNet SynSet name, and a truncated WordNet definition for the inferred SynSet.

3.4.7 Inferring Word Sense

After extracting the graph structure, our system infers a confidence distribution over word senses for each applicable node in the SpEAR graph, ignoring some pronouns, prepositions, determiners, and logical connectives. Figure 7 illustrates the output of word sense disambiguation from our system, listing all inferred word senses with a confidence score greater than 0.5. We do not interpret the highest-confidence word sense as the single "correct" word sense; rather, we regard each node as having a weighted semantic locale within a lexical ontology.

Word senses are inferred using the LMMS framework (Loureiro & Jorge, 2019): a transformer-based encoder encodes a vector for each token of the sentence. Vectors for SpEAR nodes are computed by averaging the one or more constituent token vectors. The system then computes the dot-product of each node's vector against pre-computed vectors for each word sense within its sense embeddings. The dot-product results are utilized as confidence scores.

The word sense embeddings are drawn from the SynSets (i.e., synonym sets) of WordNet, a large knowledge base containing over 117,000 word senses (Fellbaum, 2010). Computing a confidence distribution of WordNet word senses localizes each SpEAR node within a structured semantic hierarchy. This ultimately facilitates similarity-based reasoning within and across SpEAR graphs, e.g., by computing the least common ancestor between two different nodes within the WordNet semantic hierarchy. These word senses are not evaluated in this paper due to lack of ground truth WordNet labels for our datasets, but word sense disambiguation is an important cognitive capability for natural language understanding, and is facilitated by the same transformer-based NLP as the rest of the architecture.

	Dimension	**P**	**R**	**F1**	**Support**
Entities	factor	93.05	90.68	91.85	2756
	evidence	92.17	92.00	92.04	230
	epistemic	91.57	73.04	81.09	299
	association	94.60	86.83	90.54	1290
	magnitude	88.19	86.76	87.46	613
	qualifier	78.21	78.75	78.41	360
	Micro-Averaged	91.56	87.71	89.59	
Attributes	causation	44.64	68.00	53.85	342
	comparison	92.47	77.87	84.49	329
	indicates	85.38	70.00	76.73	84
	sign+	97.98	86.97	92.13	542
	sign-	90.22	72.14	80.13	202
	correlation	100.00	83.73	91.14	320
	test	–	–	–	25
	Micro-Averaged	93.85	81.23	87.08	
Relations	arg0	84.84	75.84	80.08	1325
	arg1	84.74	76.69	80.50	1384
	comp_to	77.92	59.20	67.27	187
	modifier	80.73	74.67	77.57	1582
	subtype	43.33	33.33	37.33	156
	q+	72.04	68.73	70.32	504
	q-	75.94	54.00	62.50	208
	Micro-Averaged	81.37	73.37	77.16	

Table 1: Precision, recall, F1 and support (i.e., occurrences in dataset) for SpEAR on the SciClaim dataset, using 100 held-out examples from the total 901 examples in the dataset.

4. Results

We describe two different results of using SpEAR with our qualitative causal schemata: (1) precision, recall, and F1 measure in the SciClaim scientific claims domain and (2) traversal through an ethnographic qualitative causal model. This provides empirical evidence of the effectiveness of our approach and the expressiveness of the qualitative causal schema, respectively.

4.1 Information Extraction for Scientific Claims

The SciClaim dataset for the scientific claims domain consists of 901 examples from Social and Behavior Science (SBS) literature and abstracts from PubMed and the CORD-19 dataset (Wang et al., 2020). Each example consists of a single sentence labeled by a trained NLP expert with one or more spans (possibly nested) identified as entities, zero or more attributes on each entity, and zero or more relations over entities pairs (label counts are listed in Table 1 *support*). Most datasets for transformer-based information extraction are an order of magnitude larger.

When applying SpEAR to the SciClaim dataset of scientific claims, we use the fine-tuned SciBERT transformer variant Beltagy et al. (2019) as the SpEAR input layer.

We partitioned the SciClaim dataset into a randomized split of 100 test examples and 801 training examples, and we averaged our results over 5 train/test evaluation trials. In each trial, we trained our SpEAR model for 20 epochs and then ran our evaluation. The per-class evaluations are listed in Table 1, divided across the various entities, attributes, and relations. Table 1 reports the micro-averaged results for entities, attributes, and relations, as well as support numbers to show the cardinality of each element in the full 901-example SciClaim dataset. Despite the relatively small size of the SciClaim dataset, the model achieves promising results on most classes. Our random train-test split included no examples of the **Test** attribute (which describe mentions of "ANOVA," "t-test," and other experimental methods), so Table 1 contains no results for that row.

Importantly, the relations and attributes cannot be correct if the entities they are defined over are incorrect. This means that we expect relations and attributes to have lower precision and recall, all else being equal. This is especially the case for relations, which require *both* of their constituent entities (i.e., head and tail nodes) to be properly characterized in order to be scored as correct. The relations **q+** and **q-** achieved relatively low performance, due in part to the lower support in the training data, and also due to the often greater distance between these spans in the text, all else being equal.

These results across entities, attributes, and relations support our claim that qualitative causal structure can be characterized by context-sensitive NLP models.

4.2 Extracting and Traversing Ethnography-Derived Causal Models

In the ethnographic domain, we trained SpEAR on labeled examples from Anthropology papers, ethnographic manuscripts, and tweets, all related to the topic of maternal and child health in western African countries. We then ran SpEAR to extract information from these and other sentences from the same types of literature.

This ethnographic dataset is roughly half the size of the SciClaim dataset, so rather than provide another table of F1 scores, we demonstrate the reasoning capabilities supported by the NLP-

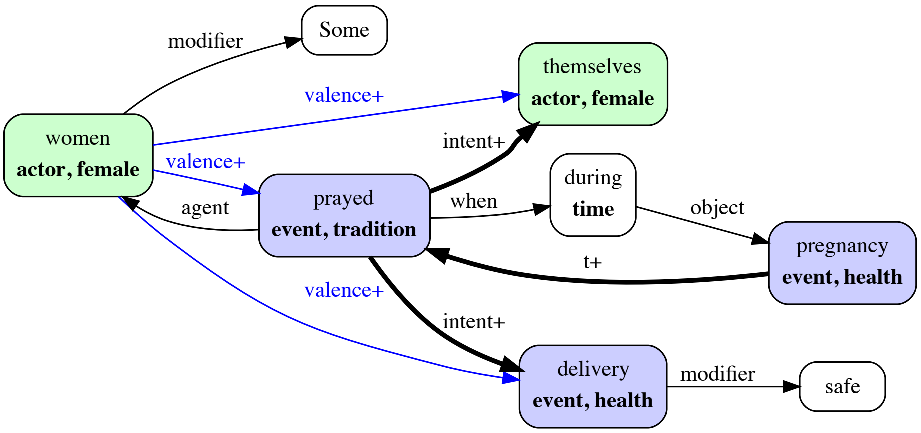

Figure 8: Knowledge graph in the ethnography schema with valence inferred, for the text "Some of the women prayed for themselves during pregnancy for safe delivery."

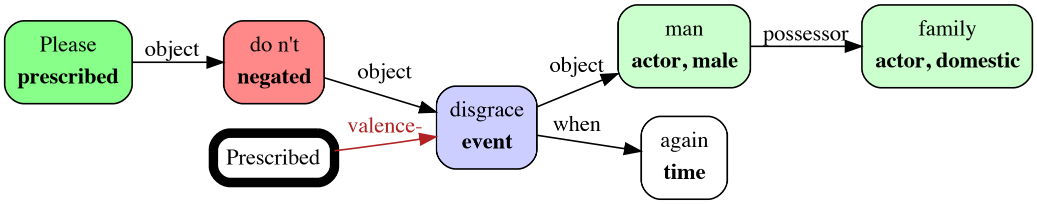

Figure 9: Knowledge graph in the ethnography schema with valence inferred, for the text "Please don't disgrace the man of the family again."

extracted causal structure. The preliminary results in Section 4.2.1 and Section 4.2.2 include both SpEAR-extracted and human-labeled (i.e., ground truth) data, so we consider this a proof-of-concept study of the ability to reason over the causal models extracted by SpEAR.

4.2.1 Computing Valence via Intentions and Qualitative Monotonicity

In the ethnographic domain, the causal models in our schema (and therefore extracted by our model SpEAR) include intention (**intent+**) and function (**function+**) relations, which indicate an agent-based desire or normative effect of an action or artifact.

These relations indicate an agent-based or normative *valence* (i.e., the positive or negative desirability) of an agent to achieve (or maximize) or prevent (or minimize) an event or quantity. Our graph schema also includes a **prescribed** attribute for occurrences of "should" and "must" and "please" that indicate a request or positive valence on behalf of the author or speaker, and a **negated** attribute for occurrences of "don't" and "not" and other negations to indicate a negation of the node's reachable subgraph. These intentional, functional, prescribed, and negated structure in the graph support graph-based inference of agents' valence.

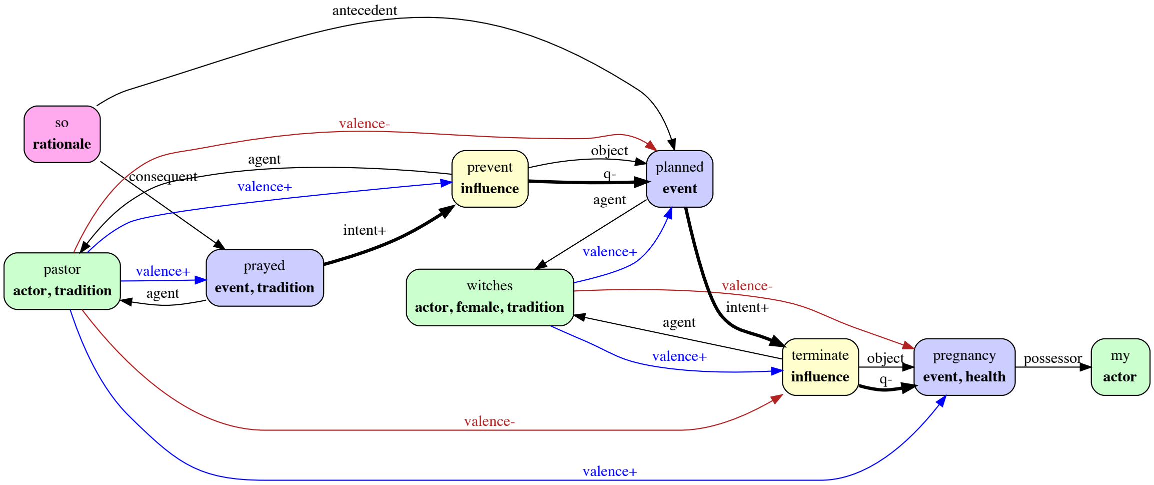

Figure 10: Knowledge graph in the ethnography schema with valence inferred, for the text "...the witches had planned to terminate my pregnancy, so the pastor prayed to prevent it."

Valence computation starts at any **intent+** or **prescribed** source node, and then traverses forward, asserting that the **agent** (or otherwise a generic **prescribed** element) has a positive valence for that node. When a **negated** attribute or **q-** relation is traversed, the sign of the valence is inverted.

In Figure 8, the traversal computes that women have positive valence for "prayed" (their direct action) and "themselves" (which they pray for) and "safe delivery" (which they also pray for). In Figure 9, the speaker prescribes the negation of an event, so the traversal asserts a normative negative valence for the disgracing the man of the family. Figure 10 is the most complex of these SpEAR semantic graphs, describing a pastor praying to prevent some witches' plan to terminate a pregnancy. The traversal infers that the pastor and the witches have opposite valences toward the plan, the pregnancy termination, and the pregnancy itself.

These simple traversals over complex SpEAR graphs can help us infer the norms and heterogeneous values of actors across cultures, from unstructured text.

4.2.2 Finding Contextual Associations in Ethnographic Causal Models

In addition to analyzing the ethnographic models on a per-sentence basis (see above), we assemble them into a global graph comprising the SpEAR graphs from each sentence from the ethnographic corpus. We implemented a simple query function that records the semantic paths between a *start* and *end* patterns, where the pattern could bind a node's lemma (e.g., "baby") or bind multi-node semantic graph structure, e.g., any **Event** with lemma "eat" whose **agent** has lemma "woman." The SpEAR nodes traversed from the query (from start to end patterns) comprise the set of relevant causal factors and relationships.

In the case of the Figure 11 traversal, the query begins at any "eat"-lemma event performed by a *mother* or *woman,* and terminates at a "baby"-lemma node. Intuitively, this queries how the mother's eating might affect a baby, and Figure 11 iterates through eating "sugarcane" **q+** to a baby's

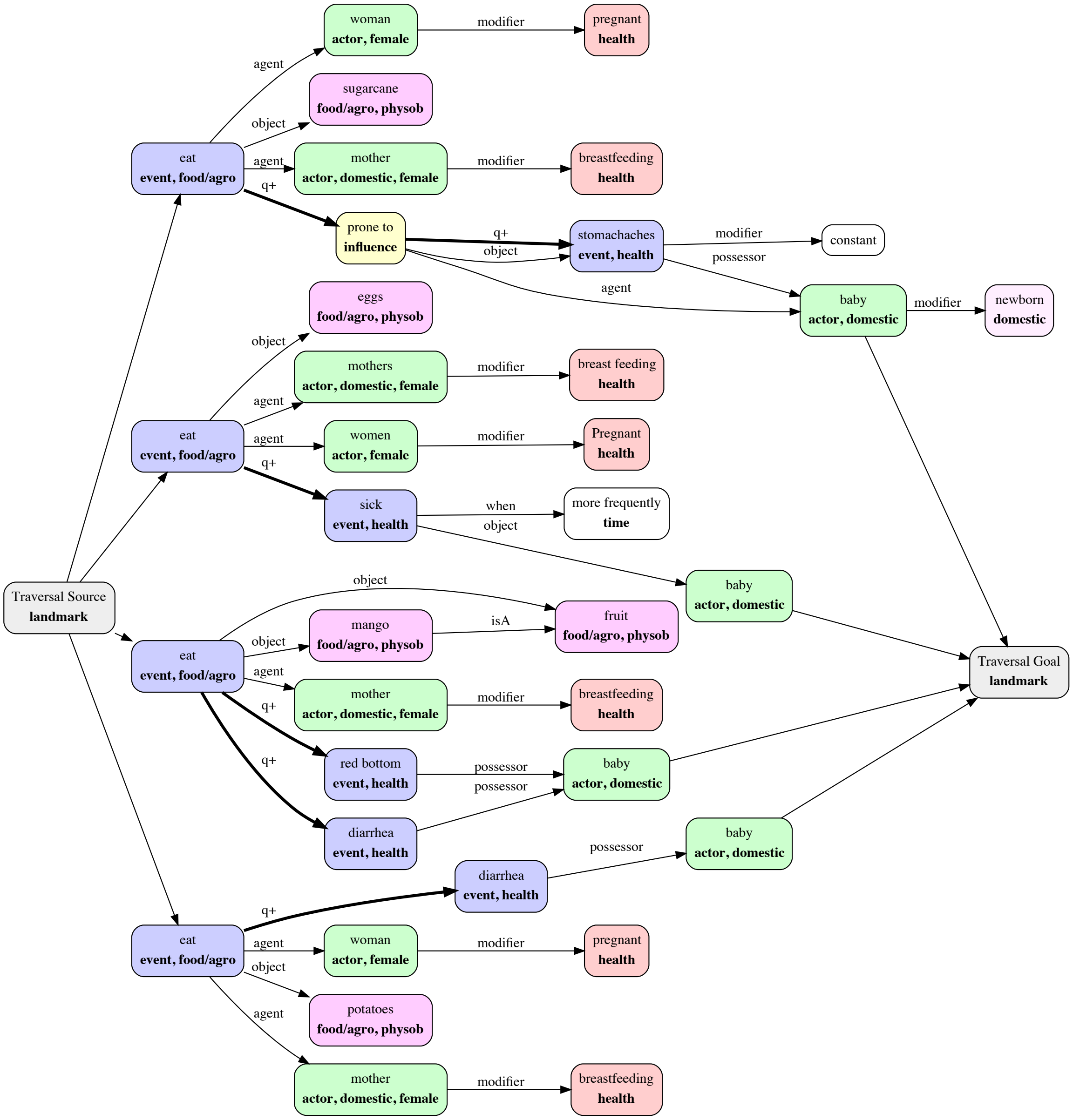

Figure 11: A graph traversal from the concept "eat" with an agent of lemma "woman" or "mother" to the concept "baby" after parsing a manuscript listing common myths about maternal and child health.

"stomachaches," through eating "eggs" **q+** to a baby being "sick," through eating "mango" **q+** to both "red bottom" and "diarrhea," etc. Note that the dietary effects extracted and traversed here are not supported by scientific evidence; rather, they are common beliefs in the regions described in the ethnography. This query-driven traversal capability provides further evidence that the SpEAR causal models support practical qualitative causal reasoning.

5. Conclusion

This paper describes our SpEAR transformer-based NLP model for extracting entities, attributes, and relationships that describe qualitative causal structure. We demonstrated the approach in the domains of (1) the SciClaim dataset of scientific claims and (2) ethnographic corpora. Our datasets are still under development, but despite their relative sparsity they support encouraging results with respect to F1-measure and practical reasoning capabilities via graph traversal.

One limitation of this work is that not all of the nodes generated by our approach are formally represented to support qualitative and numerical model-based reasoning. This is due in part to the ambiguity and hedging that we see in causal language: "smoking is associated with increased risk of lung disease" does not unambiguously specify whether we should model "smoking" as a frequency, likelihood, or single occurrence, nor does it unambiguously specify whether the risk of disease increases in likelihood or severity. The incompleteness in language—and the resulting gaps in knowledge representation—mean that some assumptions about the arguments to **q+** and **q-** may not hold in SpEAR's output: **q+** may be expressed over quantities, over events, over adjectives, or any heterogeneous mix of these, and a downstream reasoner has no formal *a priori* indicator of which these are. One remedy to this is for the NLP model to infer whether nodes are amounts, frequencies, likelihoods, etc., but it's an empirical question whether transformer-based NLP model can accurately infer these abstract categories, and it's not clear whether NLP models should attempt such inferences when the author has left it ambiguous.

Our graph traversal results suggest that the present level of representation may be adequate for use cases involving causal reasoning, graph propagation, and inferring agents' direct and indirect goals and intentions. These are important considerations for cognitive systems that reason in scientific, causal, or social domains.

As with many modern NLP architectures, the work presented in this paper utilizes a pre-trained transformer model within its architecture. Pre-trained transformers are trained on massive corpora collected from across the internet and other sources, which speeds up subsequent machine learning, but it also means that the sub-optimal biases of the training data—including racial, ethnic, gender, and other biases—become part of the models themselves. Systematic biases in pre-trained models have been well-characterized (Garg et al., 2018; Friedman et al., 2019), as have methods for debiasing them (Bolukbasi et al., 2016); however, we note that sub-optimal biases remain a risk for any machine-learned model trained on real-world text that itself contains implicit biases.

Our near-term future work is to expand our ethnographic dataset and to utilize SpEAR's results in downstream systems, e.g., for estimating the reproducibility of a scientific claim, automatically organizing and combining insights from academic literature, and globally traversing descriptive mental models to identify culture-specific, causally-potent concepts and purposes.

Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and Army Research Office (ARO) under Contract No. W911NF-20-C-0002 and W911NF-21-C-0007-04. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced

Research Projects Agency (DARPA) and Army Research Office (ARO). We thank the reviewers for their helpful feedback.

References

Alipourfard, N., et al. (2021). Systematizing Confidence in Open Research and Evidence (SCORE). *SocArXiv*.

Allen, J., de Beaumont, W., Galescu, L., & Teng, C. M. (2015). *Complex event extraction using drum*. Technical report, Florida Institute for Human and Machine Cognition Pensacola United States.

Aziato, L., Odai, P. N., & Omenyo, C. N. (2016). Religious beliefs and practices in pregnancy and labour: an inductive qualitative study among post-partum women in ghana. *BMC pregnancy and childbirth*, *16*, 1–10.

Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*, 4349–4357.

Das, D., Schneider, N., Chen, D., & Smith, N. A. (2010). Probabilistic frame-semantic parsing. *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics* (pp. 948–956).

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. From https://www.aclweb.org/anthology/N19-1423.

Eberts, M., & Ulges, A. (2020). Span-based joint entity and relation extraction with transformer pre-training. *24th European Conference on Artificial Intelligence*.

Fellbaum, C. (2010). Wordnet. In *Theory and applications of ontology: computer applications*, 231–243. Springer.

Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence*, *24*, 85–168.

Forbus, K. D. (2019). *Qualitative representations: How people reason and learn about the continuous world*. MIT Press.

Friedman, S., Schmer-Galunder, S., Chen, A., & Rye, J. (2019). Relating word embedding gender biases to gender gaps: A cross-cultural analysis. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (pp. 18–24).

Friedman, S. E., Magnusson, I. H., Schmer-Galunder, S. M., Wheelock, R., Gottlieb, J., Patel, P., & Miller, C. (2021). Toward Transformer-Based NLP for Extracting Psychosocial Indicators of

Moral Disengagement. *Annual Meeting of the Cognitive Science Community (CogSci)*.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*, E3635–E3644.

Gelman, B., Clark, C., Friedman, S. E., Kuter, U., & Gentile, J. E. (2021). Toward a robust method for understanding the replicability of research. *AAAI Workshop on Scientific Document Understanding*.

Kuipers, B. (1986). Qualitative simulation. *Artificial Intelligence*, *29*, 289–338.

Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end neural coreference resolution. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 188–197). Copenhagen, Denmark: Association for Computational Linguistics. From https://www.aclweb.org/anthology/D17-1018.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*, 167–204.

Loureiro, D., & Jorge, A. (2019). Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5682–5691). Florence, Italy: Association for Computational Linguistics. From https://www.aclweb.org/anthology/P19-1569.

Magnusson, I. H., & Friedman, S. E. (2021). Extracting fine-grained knowledge graphs of scientific claims: Dataset and transformer-based results. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mueller, R., & Abdullaev, S. (2019). Deepcause: Hypothesis extraction from information systems papers with deep learning for theory ontology learning. *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Pustejovsky, J. (1991). The syntax of event structure. *Cognition*, *41*, 47–81.

Wang, L. L., et al. (2020). CORD-19: The COVID-19 Open Research Dataset.