Extracting Qualitative Causal Structure with Transformer-Based NLP

Scott E. Friedman¹* and Ian H. Magnusson¹,² and Sonja M. Schmer-Galunder¹
¹SIFT, Minneapolis, MN, USA
²Northeastern University, MA, USA
{sfriedman, imagusson, sgalunder}@sift.net

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). This paper presents a transformer-based NLP architecture that jointly identifies and extracts (1) variables or factors described in language, (2) qualitative causal relationships over these variables, and (3) qualifiers and magnitudes that constrain these causal relationships. We demonstrate this approach and include promising results from in two use cases, processing textual inputs from academic publications, news articles, and social media.

1 Introduction

We express qualitative causal relationships in our everyday language and our scientific texts to capture the relationship between quantities or entities or events, compactly communicating how one event or purpose or quantity might be affected by another. These causal relations are not complete mechanism descriptions 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 NLP 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) qualitative causal relationships describing mechanisms, purposes, monotonicity, and temporal priority, and (3) multi-label attributes to further characterize the causal structure. Our primary claim is that context-sensitive language models can detect and characterize the qualitative causal structure of everyday and scientific language. As evidence, we present our SpEAR transformer-based NLP model based on BERT [Devlin et al., 2019] and SpERT [Eberts and 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 causal, semantic graphs produced by SpEAR do not conform to a strict ontology and therefore do not presently support the formal reasoning afforded by ontologies with strict selectional restrictions and relational constraints; however, we demonstrate that these outputs allow 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 and 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 proportionality describes how one quantity impacts another, in a directional, monotonic fashion. In this work, we designate \(\langle a, q+, b \rangle\) (and respectively, \(\langle a, q-, b \rangle\)) as qualitative proportionality from \(a\) to \(b\), such that increasing \(a\) would increase (and respectively, decrease) \(b\). This is motivated by formal quantity-to-quantity \(\alpha_{Q+/−}\) relations [Forbus, 1984] and \(M^{+/−}\) relations in qualitative simulation [Kuipers, 1986], but 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 that 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 and Carey, 2006] has acknowledged intentional (i.e., psychological, goal-based) and teleological (i.e., functional, 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 ⟨a, forPurpose, b⟩, 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 ⟨a, hasFunction, b⟩ 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 subwords) 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-the-art methodology for NLP, and has been used to process text to extract knowledge graphs, e.g., of people and relations [Eberts and Ulges, 2020]. The architecture we present in this paper has been applied to datasets of scientific claims [Magnusson and Friedman, 2021] 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.

Previous symbolic semantic parsers extract scientific claims and assertions from text with explicit relational knowledge representations [Allen et al., 2015], but many of these approaches rely on rule-based engines with hand tuning, which requires NLP experts to maintain and adapt to new domains. By contrast, our approach is designed to extract causal relationships at a level of expressiveness comparable to these symbolic systems while using the advances in transformer-based models such as SpERT [Eberts and Ulges, 2020] to learn graph-based representation from examples alone.

Other NLP approaches use machine learning to extract features from scientific texts, e.g., to identify factors and directions of influence [Mueller and Abdullaev, 2019]; however these approaches do not explicitly infer the relations between elements in the claim, as shown with the present 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 include 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 1-5 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.

**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 3.5. In Figures 1-5, attributes are rendered as parenthetical labels inside the nodes, e.g., Correlation and Increases 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 Increases 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.” Without these relations, the semantic structure of this scientific hypothesis would be ambiguous. Note that in Figures 1-5 the unlabeled arrows are all modifier relations, left blank to avoid clutter.

### 3.2 Problem Definition

We define the multi-attribute knowledge graph extraction task as follows: for a text passage $S$ of $n$ tokens $s_1, ..., s_n$, and a graph schema of entity types $T_e$, attribute types $T_a$, and relation types $T_r$, predict:

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

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

### 3.3 Knowledge Graph Schemata

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

**Scientific Claims.** Our 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 5.

This schema includes six entity types: Factors are variables that are tested or asserted within a claim (e.g., “smok-
ing 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 5; Increases expresses high or increased factor value; Decreases 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, e.g., all leftward arrows in Figures 1 through 5 (unlabeled for simplicity); 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.

This schema includes attributes for Spirituality (e.g., “God” and “prayed” in Figure 4), Action/event (e.g., “prayed” and “free” in Figure 4), Influence for causally-potent elements (e.g., “prevent” in Figure 3), and others.

We include additional relations “agent/poss” to describe the actor or possessor of an element, temporal precedence t+ relations to indicate one event preceding another, and intentional forPurpose and functional hasFunction relations to indicate the goal (i.e., intention or function, respectively) of an action or artifact.

These relatively simple statements in Figure 3 and Figure 4 originate from an ethnographic article [Aziato et al., 2016] that includes interview snippets. Despite their simplicity, the SpEAR knowledge graphs illustrate rich multi-step causality: Figure 3 indicates that prayer has the purpose of reducing the incidence (or severity of) complications, and Figure 4 plots a similar structure for praying for the purpose of preventing pain of the speaker.

3.4 Scientific Claims Dataset

Our preliminary dataset for the scientific claims domain consists of 515 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—and a larger version of our dataset will be validated in the near future—but despite the sparse dataset, our model achieves favorable performance.

3.5 Model Architecture

Our SpEAR model architecture extends SpERT with an attribute classifier. 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 pretrained values) are all trained end-to-end on our dataset.

The tokens s1, ..., sn of the text passage S are each embedded by BERT [Devlin et al., 2019] as a sequence e1, ..., en of high-dimensional vectors representing the token and its context. BERT also provides an additional “[CLS]” vector output, e0, designed to represent information from the complete text input. For all possible spans, spanj,k = s j, ..., s k, up to a given length, the word vectors associated with a span, e j, ..., e k, are combined by maxpooling to produce a single vector, e(spanj,k), where each element contains the maximum value across the token vectors for that dimension. The final span representation, x(spanj,k) is made by concatenating together e(spanj,k) and e0 along with a width embedding, w l, that encodes the number of words, l, in spanj,k. Each valid span length l looks up a different vector of learned parameters, w l.
The span representation \( x(\text{span}_{j,k}) \) is classified into mutually-exclusive entity types by a linear classifier (Figure 6a). Only spans identified as entities move on to further analysis (Figure 6b). All pairings of the remaining entities are classified for relations by a multi-label linear classifier (Figure 6c), where pairs are represented by the concatenated vectors of the two spans with the \("[CLS]"\) context vector replaced by the maxpool of the token vectors between the entities.

We implemented the component in (Figure 6d) to infer multi-label attributes on the identified entities using \( x(\text{span}_{j,k}) \) as input to another multi-label linear classifier. We take only identified entity spans as input to the attribute classifier, as this approach provided best performance and aligns with the finding by Eberts and Ulges (2020) that training on downstream tasks is best done on strong negative samples.

We partitioned our dataset into a randomized 90% train/test split of 464 and 51 examples, respectively. 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 relationships, as well as support scores to show how many examples of each element are in the full 515-example dataset. Despite the small size of our preliminary dataset, the model achieves promising results on most classes.

| Dimension | P   | R   | F1  | Support |
|-----------|-----|-----|-----|---------|
| Entities  |     |     |     |         |
| factor    | 90.13 | 86.71 | 88.39 | 1,604 |
| evidence  | 72.73 | 80.00 | 76.19 | 139   |
| epistemic | 93.33 | 100.00 | 96.55 | 178   |
| association| 95.89 | 93.33 | 94.59 | 837   |
| magnitude | 94.44 | 94.44 | 94.44 | 415   |
| qualifier | 86.96 | 68.97 | 76.92 | 216   |
| Micro-Averaged | 91.29 | 87.89 | 89.56 |        |
| Attributes |     |     |     |         |
| causation | 88.24 | 93.75 | 90.91 | 204   |
| comparison| 79.17 | 90.48 | 84.44 | 234   |
| indicates | 80.00 | 66.67 | 72.73 | 44    |
| increases | 75.86 | 95.65 | 84.62 | 262   |
| decreases | 100.00 | 100.00 | 100.00 | 134   |
| correlation| 94.74 | 94.74 | 94.74 | 199   |
| test      | 100.00 | 66.67 | 80.00 | 24    |
| Micro-Averaged | 84.62 | 91.67 | 88.00 |        |
| Relations  |     |     |     |         |
| arg0      | 82.93 | 76.40 | 79.53 | 865   |
| arg1      | 76.71 | 71.79 | 74.17 | 883   |
| comp_to   | 81.82 | 69.23 | 75.00 | 137   |
| modifier  | 84.78 | 74.29 | 79.19 | 1,080 |
| q+        | 77.78 | 56.00 | 65.12 | 295   |
| q-        | 60.00 | 85.71 | 70.59 | 138   |
| subtype   | 85.71 | 75.00 | 80.00 | 106   |
| Micro-Averaged | 81.00 | 72.97 | 76.78 |        |

Table 1: Precision, recall, F1 and support (i.e., occurrences in dataset) for each label on 10% held-out dataset using SpEAR with rectifier and filtering model.

4 Results

We describe two different results of using SpEAR with our qualitative causal schemata: (1) precision, recall, and F1 measure in the 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

For our scientific claims dataset, we use the fine-tuned SciBERT transformer variant [Beltagy et al., 2019] as the input layer of our architecture.

4.2 Traversing Ethnographic Causal Models

In the ethnographic domain, we trained SpEAR on labeled examples from Anthropology papers describing religious beliefs surrounding pregnancy in Ghana [Aziato et al., 2016]. We then ran SpEAR to extract information from these and other sentences from the same literature, resulting in a global causal graph comprising the disconnected knowledge graphs from each sentence from the literature. These preliminary results include the use of human-labeled training data, so we consider this a proof-of-concept study of the practicality of the causal structure.

We then built a traversal system to walk to and/or from any concept in this global causal graph, along the nodes and edges extracted by SpEAR. The source and destination concepts are given by the user, and then the system identifies all source and destination nodes by vector- or lemma-distance to the user’s inputs. Given these source and destination nodes, it identifies global paths from one concept to the other.

The traversal produces a graph such as the one in Figure 7, which shows paths from the source “pray” (including “prayed,” “prayer,” “prayers,” etc.) to the destination “pregnant” (also including “pregnancy.”) These traversals describe prayer for the purpose of preventing unfortunate consequences, ensuring safe pregnancy, and qualitatively increasing faith, hope, and confidence in delivery. These graph structures only contain complete paths, so the extraneous structure has been pruned for readability at the expense of completeness.

These results support our claim that the causal models extracted by our transformer-based NLP can support coarse-level reasoning and traversal across concepts.

5 Conclusion

This paper describes a transformer-based NLP model for extracting entities, attributes, and relationships that describe
Figure 7: A graph traversal from the concept “pray” to the concept “pregnant” after parsing an ethnography about spirituality in pregnancy in Ghana.

qualitative causal structure. We demonstrated the approach in the two different domains of scientific claims and descriptions of mental models from ethnography. Our datasets are still under development, but despite their relative sparsity they support encouraging results with respect to F1-measure and traversal.

One limitation of this work is that the nodes are not grounded in a formal hierarchical ontology. This means that many of the 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 to use our node-based attributes to express these different types of elements, but whether the transformer-based NLP model can accurately classify these abstract categories is an empirical question.

Our near-term future work is to expand our datasets and utilizing 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 culturally-specific causally-potent concepts and purposes.

References

[Alipourfard et al., 2021] Nazanin Alipourfard, Beatrix Arendt, Daniel Jacob Benjamin, Noam Benkler, Michael Bishop, Mark Burstein, Martin Bush, James Caverlee, Yiling Chen, Chae Clark, et al. Systematizing confidence in open research and evidence (score). 2021.

[Allen et al., 2015] James Allen, Will de Beaumont, Lucian Galescu, and Choh M Teng. Complex event extraction using drum. Technical report, Florida Institute for Human and Machine Cognition Pensacola United States, 2015.

[Aziato et al., 2016] Lydia Aziato, Philippa NA Odai, and Cephas N Omenyo. Religious beliefs and practices in pregnancy and labour: an inductive qualitative study among post-partum women in ghana. BMC pregnancy and childbirth, 16(1):1–10, 2016.
[Beltagy et al., 2019] Iz Beltagy, Kyle Lo, and Arman Co- 
hahn. Scibert: A pretrained language model for scientific 
text. arXiv preprint arXiv:1903.10676, 2019.

[Das et al., 2010] Dipanjan Das, Nathan Schneider, Desai 
Chen, and Noah A Smith. Probabilistic frame-semantic 
parsing. In Human language technologies: The 2010 an-
nual conference of the North American chapter of the 
association for computational linguistics, pages 948–956, 
2010.

[Dennett, 1989] Daniel Clement Dennett. The intentional 
stance. MIT press, 1989.

[Devlin et al., 2019] Jacob Devlin, Ming-Wei Chang, Ken-
ton Lee, and Kristina Toutanova. BERT: Pre-training of 
deep bidirectional transformers for language understand-
ing. In Proceedings of NAACL-HLT 2019, pages 4171– 
4186, Minneapolis, Minnesota, June 2019. Association for 
Computational Linguistics.

[Eberts and Ulges, 2020] Markus Eberts and Adrian Ulges. 
Span-based joint entity and relation extraction with trans-
former pre-training. 24th European Conference on Artifi-
cial Intelligence, 2020.

[Forbus, 1984] Kenneth D Forbus. Qualitative process the-
ory. Artificial intelligence, 24(1-3):85–168, 1984.

[Forbus, 2019] Kenneth D Forbus. Qualitative representa-
tions: How people reason and learn about the continuous 
world. MIT Press, 2019.

[Friedman et al., 2021] Scott E. Friedman, Ian H. Magnus-
son, Sonja M. Schmer-Galunder, Ruta Wheelock, Jeremy 
Gottlieb, Pooja Patel, and Christopher Miller. Toward 
Transformer-Based NLP for Extracting Psychosocial In-
dicators of Moral Disengagement. In CogSci, 2021.

[Gelman et al., 2021] Ben Gelman, Chae Clark, Scott E 
Friedman, Ugur Kuter, and James E Gentile. Toward a 
robust method for understanding the replicability of re-
search. In AAAI Workshop on Scientific Document Un-
derstanding, 2021.

[Kuipers, 1986] Benjamin Kuipers. Qualitative simulation. 
Artificial intelligence, 29(3):289–338, 1986.

[Lombrozo and Carey, 2006] Tania Lombrozo and Susan 
Carey. Functional explanation and the function of expla-
nation. Cognition, 99(2):167–204, 2006.

[Magnusson and Friedman, 2021] Ian H. Magnusson and 
Scott E. Friedman. Graph knowledge extraction of causal, 
comparative, predictive, and proportional associations in 
scientific claims with a transformer-based model. In AAAI 
Workshop on Scientific Document Understanding, 2021.

[Mueller and Abdullaev, 2019] Roland Mueller and Sardor 
Abdullaev. Deepcause: Hypothesis extraction from inform-
ation systems papers with deep learning for theory on-
tology learning. In Proceedings of the 52nd Hawaii Inter-
national Conference on System Sciences, 2019.

[Pustejovsky, 1991] James Pustejovsky. The syntax of event 
structure. Cognition, 41(1-3):47–81, 1991.

[Wang et al., 2020] Lucy Lu Wang, Kyle Lo, Yoganand 
Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Bur-
dick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney 
Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul 
Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, 
Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan 
Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, 
Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Se-
bastian Kohlmeier. Cord-19: The covid-19 open research 
dataset, 2020.