Combining Extraction and Generation for Constructing Belief-Consequence Causal Links

Maria Alexeeva
Department of Linguistics
University of Arizona
Tucson, AZ
alexeeva@email.arizona.edu

Allegra A. Beal Cohen
Agricultural and Biological Engineering Department
University of Florida
Gainesville, FL
aa.cohen@ufl.edu

Mihai Surdeanu
Computer Science Department
University of Arizona
Tucson, AZ
surdeanu@email.arizona.edu

Abstract

In this paper, we introduce and justify a new task—causal link extraction based on beliefs—and do a qualitative analysis of the ability of a large language model—InstructGPT-3—to generate implicit consequences of beliefs. With the language model-generated consequences being promising, but not consistent, we propose directions of future work, including data collection, explicit consequence extraction using rule-based and language modeling-based approaches, and using explicitly stated consequences of beliefs to fine-tune or prompt the language model to produce outputs suitable for the task.

1 Introduction

Natural language processing can successfully capture the causal dynamics present in many complex systems. This type of automated extraction is particularly useful for computational modelers, who may be faced with a large and complex domain literature that cannot be easily summarized by humans. Information extraction systems like Eidos (Sharp et al., 2019) can help modelers build skeleton models of causes and effects present in systems by extracting causal links that exist between entities and processes.

While many causal dynamics are mechanistic, such as water level driving crop yield, other dynamics are driven by subjective factors, such as the political beliefs of a population driving their decisions to wear masks. Extracting these dynamics comes with two challenges: Extracting the beliefs and consequences present in the text, and inferring implicit consequences of beliefs. For example, the following sentence contains both a belief and an explicit consequence:

1. *Peanut and maize are generally sown after a few big rains when farmers believe that the rainy season has really started.*

The above sentence can be represented by a binary, directed causal link, where the first node is the belief about the rainy season and the second node is the consequence of the belief (crop sowing). However, the consequences of beliefs are frequently implied, such as in the following sentence:

2. *Also use of chemicals and machinery on their paddy field is often considered undesirable.*

To a human, the obvious consequence is that the farmers will not use chemicals, but the text does not explicitly state this. A modeler wants to generate causal belief-consequence pairs from a large literature without annotating every implicit consequence; thus, methods of automating belief extraction ought to account for implicit consequences.

In this paper, we address the problem of extraction of beliefs and their consequences with a novel extraction + generation approach. We first extract beliefs using an event extraction grammar; and we then use text generation with large language models (LM) to generate possible consequences of the extracted beliefs when no consequence is stated in text. We expect that given a belief and its context, there is only a limited number of possible consequences humans can infer. For the consequence generation approach to be considered successful, we want machine-generated consequences to match those produced by humans—that would be an indicator that generated beliefs are indeed relevant for the model.

With this work, we make the following contributions:

- We define a new task—causal link extraction based on beliefs—which can be used to enrich models with subjective beliefs of local populations.
- We conduct a qualitative analysis of automatic consequence generation. We find that InstructGPT-3 model (Ouyang et al., 2022),

which we use, is able to produce consequences relevant to beliefs, but does not seem to make consistently relevant predictions.

- We propose the next steps for this project, which include collecting and annotating data for the task, explicit consequence extraction, and using explicitly stated consequences for fine-tuning or prompting language models to make their outputs consistently relevant for the task.

2 Related Work

2.1 Modeling causality.

Causality modeling is a popular area of investigation thanks to its usefulness for multiple applications, e.g., question answering (Sharp et al., 2016). Both rule-based approaches (e.g., Sharp et al., 2019) and deep learning approaches (Li et al., 2021) have been proposed. We are not aware of any other work that investigates causal links rooted in beliefs.

2.2 Rule-based extraction.

Rule-based approaches have been shown to be powerful and robust, e.g., by Valenzuela-Escárcega et al. (2015) with their rule-based information extraction framework Odin. The framework allows for both surface and syntactic dependency-based rules and has been successfully used for extracting information in several projects, including protein interaction extraction (Valenzuela-Escárcega et al., 2018) and causal events extraction (Sharp et al., 2019).

2.3 Automatic text generation

Most recently, OpenAI released models that were trained to allow for human-augmented text generation, in which the user can provide the model with prompts either defining the task or providing examples to the model to demonstrate the task in a few shot setting (Ouyang et al., 2022). We use this model in our experiments.

3 Procedure

We automatically extracted beliefs from a collection of documents—scientific publications and reports—related to agriculture and social norms of Senegal. We then double-annotated fifty of those beliefs with whether or not their consequences were explicitly stated in one-sentence and one-paragraph context windows. When there was no explicit consequence stated, the annotators provided the consequences they believed to be fitting based on the belief and one paragraph-long context. We also compared human-generated implicit consequences with those generated by the InstructGPT-3 model (*text-davinci-001* in the API) (Ouyang et al., 2022).

```
- name: belief-rule
  label: Belief
  type: dependency
  pattern: |
    trigger = [lemma=/consider/]
    believer:Agent = /nsubj/
    belief:Proposition = /xcomp/
```

Figure 1: A sample rule for extracting beliefs implemented using the Odin information extraction framework (Valenzuela-Escárcega et al., 2015)

3.1 Belief and Explicit Consequence Extraction

For extracting beliefs, we converted PDF files to text files using the pdfminer.six package and used the Odin rule-based information extraction framework (Valenzuela-Escárcega et al., 2015) for extraction. Using the framework, we wrote a grammar based on a set of triggers indicating beliefs, e.g., *think*, *believe*, *consider*, etc, and extracted events with believer (optional) and belief arguments. A sample rule is in Figure 1. We excluded beliefs of the author of the documents and only extracted reported beliefs (Prabhakaran et al., 2015)—in our case those are the beliefs of the local population.

Explicitly-stated consequences can be extracted using a rule-based approach like we do with beliefs. While the rule-based framework that we use supports same sentence extraction with cross-sentence coreference resolution, to extract consequences across sentences, the framework will need to be expanded. We leave the task of extracting explicit consequences to future work.

3.2 Implicit Consequence Generation

For the beliefs that are not accompanied by explicit consequences, we generated consequences using the InstructGPT-3 model (Ouyang et al., 2022). We primed the model with six few-shot examples with the following structure: "Belief: <text of belief extraction> Consequence: <text of a possible consequence>", e.g.:

3. **Belief:** *Rice grown in the dry season produced higher yields and was perceived to have lower risks.*
   Consequence: *Farmers may not need to buy insurance for rice grown during the dry season.*

For creating the prompts, we used beliefs that were automatically extracted from text. The consequences in the prompts were either taken directly from text or were created by the authors to match the task. Both beliefs and consequences taken directly from text were edited slightly for clarity. Additionally, we experimented with providing the model with fewer examples (two and four in addition to six as discussed above) and also prompting the model to generate a consequence by using a discourse marker *That's why* without including any belief-consequence pairs as examples. We did not do any prompt tuning.

3.3 Evaluation

We do a qualitative analysis of human and machine-generated implicit consequences. For every belief, we manually inspect the two consequences produced by the human annotators and judge them to be the same if there is an overlap in context even if the form—the exact wording—is different. For automatically-generated vs. human-generated comparison, we consider the generation successful if at least one out of three automatically-generated consequences overlaps with at least one of the human-generated consequences.

Additionally, we evaluate the quality of automatically-generated consequences in terms of their relevance to the belief prompt, regardless of their similarity to human-generated consequences.

4 Results and Discussion

Based on the comparison of two sets of annotations, we see that a large number of beliefs do not have associated explicitly-stated consequences: the two annotators judged an average of 72% of the 50 beliefs annotated to not have consequences explicitly stated within the same sentence and an average of 49% to not have them within the one paragraph context window. These results indicate that both extraction and generation have to be included in the approach.

Analyzing the 18 beliefs that both annotators agreed did not have explicitly stated consequences, we see that, as expected, annotators tend to agree on possible consequences of beliefs: for 72% of beliefs, human annotators produced potential consequences with similar content (Table 1). We also see that there is promise for generating consequences using large language models: the GPT-3 model can produce consequences that match those produced by human annotators:

Condition	Overlap
two annotators	13 (72%)
GPT-3 and one annotator	12 (66%)
GPT-3 and both annotators	9 (50%)

Table 1: Overlap in content between different consequences produced (based on 18 beliefs with no consequences explicitly stated in text).

4. **Belief:** *Planners and technicians feel that the development of irrigation systems could offer a solution to the crisis in food production in Africa.*
   Annotator 1: *Planners and officials will invest more in the development of irrigation systems.*
   Annotator 2: *They should develop irrigation systems.*
   GPT-3: *Planners and technicians focus on the development of irrigation systems.*

However, while producing some consequences that overlap with those produced by human annotators (Table 1), GPT-3 also generates text that, while thematically relevant to the prompt, does not constitute a successful consequence generation. To evaluate consequence generation independently from that done by human annotators, we analyze 54 GPT-3-generated consequences (three per each of the 18 beliefs with no explicit consequences) for whether or not they are appropriate for the corresponding beliefs. We judge 40 of the GPT-3-generated consequences (74%) to be possible consequences for the given belief prompt.

The quality of several consequences generated for each belief is not necessarily consistent. As seen from Table 2, for a given belief, all, some, or none of the three generated consequences can be appropriate. This poses a potential issue for downstream tasks in how there is no way to verify that a correct prediction was generated or selected from several generated predictions. We see several ways in how this could be addressed. First, we believe that with additional training using a dedicated data

Condition	Count
all correct	8
a mix of correct and incorrect	7
all incorrect	3

Table 2: Counts of beliefs for which all three generated consequences were correct, some were correct, or none were correct.

set, consequences of beliefs can be generated more consistently. Second, following Lu et al. (2021), we could apply logical or lexical constraints on the decoding side. Third, with several consequences generated, we may be able to assign higher weights to consequences that overlap in content with the other consequences generated for the same belief. Finally, this approach can be used for augmentation, not automation of some human efforts, e.g., modeling, in which case the domain expert will be able to evaluate produced belief-consequence pairs before using them.

Some of the error types observed among the 54 consequences generated by the GPT-3 model in a few-shot setting and example sentences to illustrate the error types are in Table 3. The counts of the error types are in Table 4. We note that for this analysis, not all error types necessarily disqualify a prediction from being a possible consequence. For instance, in the following example, while the specific dollar values make us judge the generated text to be too specific, a budget increase is indeed a possible consequence of a decision to rehabilitate the land:

5. **Belief:** *[...] personnel believe that up to an additional 60,000 ha could be rehabilitated.*
   Consequence: *They recommend an increase in the annual budget for rehabilitating degraded lands from $10 million to $15 million.*

In our experiments, we also see that providing examples to the model is crucial. In a zero-shot setting (i.e., with no belief-consequence examples provided to the model), the model did generate some propositions that could be seen as appropriate consequences for given beliefs. However, in this setting, in most cases the model also continued generating text until the word limit was reached. This was not the case for the majority of consequences generated in the few-shot setting: following the format of the provided examples, the model stopped generating after one sentence. Moreover, without explicit examples provided to the model, the generated text tended to contain information that was too specific to count as a successfully generated consequence. This included people's or company names, hot political topics (see example 6), or conversation markers (see example 7).

6. **Belief with a prompt:** *[...] centers are often considered the best private family planning clinics. That's why*
   Consequence: *the Trump administration's proposed rule to cut $300 million in Title X family planning funding is so alarming. Title X*

7. **Belief with a prompt:** *people believe that their hard work in the rice paddies is part of a covenant with their supreme deity. That's why*
   Consequence: *they work so hard." "I see. So their hard work is a form of worship." "Exactly*

We do not observe major differences between providing two vs. four vs. six examples to the model.

5 Future Work

To address the issues with GPT-3 generations discussed in the paper, in future work, we want to explore the possibility of using explicitly stated consequences as data to improve explicit consequence generation by large language models, both for fine-tuning and as few-shot examples.

For extracting explicit consequences, we will be exploring expanding the current rule-based framework to cross-sentence extractions as well as using GPT-3 text generation for event extraction. With the latter, the task can be framed in the following way: given a belief and its context, the model should produce the text that matches the text span that indicates the consequence of the belief.

For both these directions, we need to have a corpus of belief-consequence annotations, which means we will be conducting an annotation exercise. This will also help us support our findings regarding distributions of explicit and implicit consequences as well as the ability of human annotators to infer implicit consequences based on the belief and its context.

	Belief	Consequence	Error Type
1.	women unanimously believe their work to be profitable and well worth pursuing.	Women continue to engage in embroidery work.	too specific
2.	Mothers also believe that a child does not need to eat solid food before 8 months.	Breastfeeding rates are low in the study area.	contradiction
3.	farmers strongly feel that increased bird pressure intensity with resulting higher bird damages are the main reason for these lower yields in 2006.	Farmers stop growing rice.	wrong focus
4.	gifts, financial support, and physical attention are themselves considered important demonstrations of romantic love.	Men in this culture often give gifts, offer financial support, and provide physical attention to more than one woman at a time.	unsubstantiated consequence
5.	personnel believe that up to an additional 60,000 ha could be rehabilitated.	The additional 60,000 ha will not be rehabilitated.	wrong polarity
6.	use of chemicals and machinery on their paddy field is often considered undesirable.	Use of chemicals and machinery on their paddy field is often considered undesirable.	restatement

Table 3: Some of the error types observed in belief consequences generated by GPT-3.

Error Type	Count
too specific	9
contradiction	3
wrong focus	2
unsubstantiated consequence	1
wrong polarity	1
restatement	1

Table 4: Some of the error types and their counts observed observed in the 54 consequences generated by GPT-3 for the 18 beliefs (three consequences generated per belief).

Finally, we want to use belief-consequence pairs to build cognitive models of decision-making, e.g., modeling how a belief about rains causing crop damage might cause the believer to harvest early.

6 Conclusion

In this paper, we introduce the task of causal link extraction based on beliefs. We propose an approach for the task that combines extraction and generation, and provide a small-scale, qualitative analysis of a large language model performance on the task. Additionally, we outline directions of future work.

Acknowledgements

The authors thank the anonymous reviewers for helpful discussion and Andrew Zupon for help with annotation.

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under the Habitus program. Maria Alexeeva and Mihai Surdeanu declare a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

References

Zhongyang Li, Xiao Ding, Kuo Liao, Bing Qin, and Ting Liu. 2021. Causalbert: Injecting causal knowledge into pre-trained models with minimal supervision. *arXiv preprint arXiv:2107.09852*.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4299, Online. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Preprint*.

Vinodkumar Prabhakaran, Tomas By, Julia Hirschberg, Owen Rambow, Samira Shaikh, Tomek Strzalkowski, Jennifer Tracey, Michael Arrigo, Rupayan Basu, Micah Clark, Adam Dalton, Mona Diab, Louise Guthrie, Anna Prokofieva, Stephanie Strassel, Gregory Werner, Yorick Wilks, and Janyce Wiebe. 2015. A new dataset and evaluation for belief/factuality. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 82–91, Denver, Colorado. Association for Computational Linguistics.

Rebecca Sharp, Adarsh Pyarelal, Benjamin Gyori, Keith Alcock, Egoitz Laparra, Marco A. Valenzuela-Escárcega, Ajay Nagesh, Vikas Yadav, John Bachman, Zheng Tang, Heather Lent, Fan Luo, Mithun Paul, Steven Bethard, Kobus Barnard, Clayton Morrison, and Mihai Surdeanu. 2019. Eidos, INDRA, & delphi: From free text to executable causal models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 42–47, Minneapolis, Minnesota. Association for Computational Linguistics.

Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. 2016. Creating causal embeddings for question answering with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 138–148, Austin, Texas. Association for Computational Linguistics.

Marco A Valenzuela-Escárcega, Özgün Babur, Gus Hahn-Powell, Dane Bell, Thomas Hicks, Enrique Noriega-Atala, Xia Wang, Mihai Surdeanu, Emek Demir, and Clayton T Morrison. 2018. Large-scale automated machine reading discovers new cancer-driving mechanisms. *Database*, 2018.

Marco A. Valenzuela-Escárcega, Gus Hahn-Powell, Mihai Surdeanu, and Thomas Hicks. 2015. A domain-independent rule-based framework for event extraction. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 127–132, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.