MAIN ARTICLE

From text to model: Leveraging natural language processing for system dynamics model development

Guido A. Veldhuis,[a,b*] Dominique Blok,[a] Maaike H.T. de Boer,[a] Gino J. Kalkman,[a] Roos M. Bakker[a,c] and Rob P.M. van Waas[a]

Abstract

Textual data is abundantly available, and natural language processing (NLP) facilitates its analysis. However, system dynamics (SD) modelling relies on the modeller to identify relevant information. We explore the ability of NLP models to support SD modelling by identifying causal sentences in texts. We provide a primer on the notion of causality in SD and on the linguistic properties of causality, followed by an introduction of NLP models suitable for this task. Using three test cases, we evaluate the performance of the NLP models using common evaluation metrics and an SD model completeness metric. We conclude that NLP models can add considerable value to SD modelling, provided that remaining challenges are addressed. One such caveat is the difference we observe between information regarded as causal and information relevant for describing system structure. We discuss how these challenges can be addressed through collaboration between the NLP and SD fields.

Syst. Dyn. Rev. (2024)

INTRODUCTION

System dynamics (SD) modelling aims to describe the causal structure and behaviour of a complex dynamic problem to propose policies that lead to lasting improvement in system behaviour. For this task SD modellers can draw on information from mental, written and numerical sources (Forrester, 1980, 1992). Written data bridges the gap between sparsely available numerical data and the rich, but hidden, world of mental data. Written data sources contain 'concepts and abstractions that interpret other information sources' (Forrester, 1980 p., 557). This includes mental models, which would otherwise remain invisible. Written data can be used throughout the SD modelling process to define the purpose, structure and parameters of a model, as well as to develop policies, perform validation tests and facilitate implementation. Today, more written data than ever are available. Businesses, governments and other sources produce volumes of text data in various forms, such as reports, minutes, articles and tweets. For example, an exponentially growing body of academic literature is accessible, with a major search engine now indexing over 389 million records (Gusenbauer, 2019). However, availability of more written data will do little to improve the application of

[a] TNO—The Netherlands Organization for Applied Scientific Research, The Hague, The Netherlands
[b] Institute for Management Research (IMR) Radboud University, Nijmegen, The Netherlands
[c] Leiden University Centre for Linguistics (LUCL), Leiden, The Netherlands
* Correspondence to:
Guido A. Veldhuis, TNO—The Netherlands Organization for Applied Scientific Research, Oude Waalsdorperweg 63, 2597 AK The Hague, The Netherlands.
E-mail: guido.veldhuis@tno.nl

Accepted by Birgit Kopainsky, Received 19 October 2023; Revised 8 March 2024 and 19 April 2024; Accepted 3 May 2024

System Dynamics Review
System Dynamics Review
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/sdr.1780

SD without the means to process this information in both a meaningful and efficient manner.

In SD modelling, written data is an indispensable source of information about the causal structure of the problem under investigation. For example, organisational documentation provides the SD modeller with insight into an organisation's operations. Academic papers explore and test causal relations. Opinion articles offer a glimpse inside the assumptions people hold about the structure and behaviour of a system. Strategic documents reveal the causal logic employed by executives. Similarly, the minutes from council meetings provide a window into how public officials interpret system behaviour and subsequently act.

Although the SD field has emphasised the importance of written data, techniques to do so have received limited attention. Some authors have developed qualitative research techniques that can be used to systematically extract meaning from written sources, focusing on inferring model structure from text (Eker & Zimmermann, 2016; Kim & Andersen, 2012; Kunc et al., 2023; Luna-Reyes & Andersen, 2003; Tomoaia-Cotisel et al., 2022; Turner et al., 2013; Yearworth & White, 2013). Their work addresses procedures, techniques and software used to analyse text and document the findings. However, these methods often involve labour-intensive tasks like identifying and annotating text fragments, with software playing a supportive rather than a central role (Eker & Zimmermann, 2016; Luna-Reyes & Andersen, 2003). For example, Yearworth and White (2013) use software to support the user to some extent by suggesting potential relations based on the proximity of coded concepts in the text. However, proximity is no guarantee for causality. Sterman (2018, p. 40) described leveraging data to 'develop, test, communicate and implement rigorous, reliable and effective insights in the dynamics of complex systems' as the 'next frontier' for SD. In this paper we will take a step towards this frontier by illustrating how natural language processing (NLP) models can be used to extract sentences that contain causal information from documents—information which can then be used to support the SD modelling process using existing techniques.

The NLP field combines computer science and linguistics to develop models that can perform a range of tasks related to the processing, interpretation and generation of text. This includes the task of identifying semantic relations in texts, such as causal relations. In NLP, causal relations are commonly defined as 'an event (cause) that results in another event (effect) to happen or hold' (Mostafazadeh et al., 2016, p. 55).[1] Causality can be expressed both implicitly and explicitly. Explicit expressions of causality often involve specific verbs, which can be divided into three categories (Nedjalkov & Silnickij, 1973): (1) simple causatives: a synonym of the verb cause, such as *generate*; (2) resultative causatives: a verb linking a cause with a resulting situation, such as *break*; (3) instrumental causatives: a verb that contains part of the event and the result, such as *clean*. A more in-depth discussion of identifying causality in text is provided in the section 'Causality in natural language'.

[1]It is important to note here that in the NLP field events are not necessarily limited to specific occurrences in time and space. As defined by Mostafazadeh et al. (2016, p. 52), drawing on Pustejovsky et al. (2003), 'An event is any situation (including a process or state) that happens, occurs or holds to be true or false during some time period (punctual) or time interval (durative).'

DOI: 10.1002/sdr

While the NLP field has long worked on identifying text fragments containing causal information, recent years have seen significant improvements in performance (Yang et al., 2022). However, causal information extraction is still considered an 'open problem' (see Akkasi & Moens, 2021, p. 1; Asghar, 2016; Feder et al., 2022; Yang et al., 2022). With the advancements in artificial intelligence in the last decades, trained transformer-based models have become an important technique for causal extraction and other NLP tasks. In this article, we will discuss and test several of such models.

We will first discuss the notion of causality in SD modelling, addressing its definition and identification in text, which has received limited attention in SD literature. This discussion leads us to how causality is expressed in natural language and criteria used in the NLP field to identify causality. We then discuss several challenges facing causal information extraction models. This is followed by a discussion of different NLP models for causal extraction. Next, we demonstrate the practical application of causal information extraction, introducing a software pipeline that preprocesses text data and employs one of five different models to extract sentences containing causal information. We compare the performance of these models through three test cases, using common NLP evaluation metrics and introducing a model completeness score to measures whether sentences referring to known model relations are detected by the NLP models. Finally, we discuss the results, limitations and the potential of NLP technology to support SD model development.

CAUSAL INFORMATION IN TEXT

Causality in SD modelling

Causality is a central concept in SD modelling, but its definition and identification have only received sporadic attention (see: Pedercini, 2006; Schaffernicht, 2010). SD modellers aim to understand and influence behaviour. To achieve this, they must piece together information to construct and test a dynamic hypothesis that describes how behaviour over time results from the causal interaction between variables.

A dynamic hypothesis is captured in a CLD or stock and flow (SF) model and describes the structure of the system responsible for the behaviour of interest. A CLD contains 'causal links', depicted as arrows, that 'denote the causal influences among the variables' (Sterman, 2000, p. 138). Various diagramming techniques from the broader field of (soft) operational research use arrows but their meaning varies (Lane & Husemann, 2010). The arrow in a CLD is defined as: 'A positive (negative) link means that if the cause increases, the effect increases (decreases) above (below) what it would otherwise have been, and if the cause decreases, the effect decreases (increases) below (above) what it would otherwise have been' (Sterman, 2000, p. 139).

For SD modellers 'the term "causal" has a specific operational connotation' (Olaya, 2016, p. 200). Through operational thinking, SD modellers seek 'intelligible explanations' (p. 201) of how a system operates. The type of causal information that SD modellers need to construct a model becomes clearer in SF models.

DOI: 10.1002/sdr

In its simplest form an SF model describes a system in one or several stocks and features two distinct types of relations: information relations and flows. Flows determine how a stock can change over time and describe where decisions and activity occur within the system. The value of a flow depends on the value of one or more stocks or constants. The relation from a stock or a constant to a flow, known as an information relation, describes how information translates into activity. On closer inspection, the SD method thus has two distinct forms of relations that are used to describe the operations of a system: flows that influence stocks and information relations that influence flows.

SD models describe causality in an aggregated form and generally do not aim to explain single events but system behaviour which '… reveals itself as a series of events' (Meadows, 2008, p. 89). Schaffernicht (2010, p. 656) observed that SD 'deals with how behaviour causes behaviour'. Thus, a definition of causality in SD terminology could read: *behaviour-over-time that results in other behaviour-over-time to occur*. Events in this context can relate to a 'specific episode of the behavior', 'a change of behavior' or 'a change [in the mode] of behavior' (Schaffernicht, 2010, p. 656). In extracting information from text, both an aggregated behaviour-over-time and event perspective are relevant. Information on aggregate causal structure and behaviour can benefit the modelling process directly. Observations of the causal relations between specific events reveal information that can lead to the identification of aggregated causal structures and behaviours. NLP models generally consider both perspectives, but they may ignore non-causal information that is relevant for SD modelling, as we will observe in our results and revisit in the 'Discussion' section.

SD modellers will analyse texts to identify the stocks, flows and information relations relevant for understanding the problem under investigation. However, the SD literature provides little guidance on how text fragments that contain useful causal information are identified. Among the few, Eker and Zimmermann (2016) give some insight into the criteria they use (p. 8): 'To establish such causal relationships, we looked out for indicators such as "because", "if … then", but we certainly also used our general understanding that someone expresses a causal relationship'. In the next section we will provide a more in-depth look at how causal information can be recognised in text.

Causality in natural language

Identifying expressions of causality in text is considerably less straightforward than might seem based on the simple definitions used in SD. This section offers a primer and addresses three challenges: the variety of ways in which causality can be expressed, inter- versus intra-sentential causality and explicit versus implicit expressions of causality.

Natural language has a variety of ways to express causality. For instance, consider the examples in (1)–(10):

1. Viruses can cause a program to stop functioning.
2. Global warming is the effect of CO_2 emissions.
3. Laura finished her PhD, so she can now apply for postdocs.

DOI: 10.1002/sdr

4. Because no one passed the exam, the whole class will have to sit the exam again.
5. Caitlin worked late on Wednesday. Therefore, she was tired on Thursday.
6. If this law is passed, then childcare will be free for children below the age of twelve.
7. If it hadn't rained on Tuesday, we would have gone swimming.
8. Daniel scrubbed the pans clean.
9. Harry sang a song. Everyone in the room was quiet.
10. Happy with the result of the negotiations, the union leaders agreed to sign the deal.

In these examples, we see that causality can be expressed through causal verbs such as *to cause* in (1), causal nouns such as *effect* in (2), adverbial markers such as *so* in (3), *because* in (4) and *therefore* in (5), conditional ('if … then') constructions in (6) and (7) and resultative constructions in (8). In addition, causality can be implied without overt markers: in (9), readers can induce that everyone in the room was quiet *because* Harry sang a song, but this is not indicated by an overt linguistic element. Similarly, in (10), we infer that the union leaders agreed to sign the deal as a result of being happy with the result of the negotiations, although there is no overt causal marker that conveys this information.

Causality can be expressed within one sentence, as seen in (1)–(4) and (6)–(8), or across multiple sentences, as demonstrated in (5) and (9). The first type is called *intra-sentential* causality and the latter type is referred to as *inter-sentential* causality. The field of NLP has thus far focused mostly on intra-sentential causality due to the increased complexity of including inter-sentential causality (but see, e.g., Jin et al., 2020, for an example of a method that tackles inter-sentential causality). In this paper we will follow the former path, leaving inter-sentential causality for future research.

Causality can be expressed *explicitly* or *implicitly*. Explicit causality can be recognised by clear indicators such as the verb *to cause* in (1), the noun *effect* in (2) or the adverb *so* in (3). Sentences (9) and (10) represent cases of implicit causality, where no causal marker is present. In these cases, we must infer causality using our knowledge. In the case of (9), we use our understanding that a person singing a song commonly leads to others being quiet. In (10), causality is inferred as we know that when someone is happy with the result of negotiations, this is likely to cause them to agree to sign the deal.

While developing an SD model we are interested in extracting all relevant causal information from a text, whether it's expressed explicitly or implicitly. Overlooking information might lead us to develop an invalid model. Different NLP techniques can be used to this end, which we will discuss in the next section.

Automatic causal information extraction

Broadly speaking, there are two types of methods that have been used in the domain of automatic causality extraction: rule-based methods and statistical or machine learning-based methods. Initially, rule-based methods were commonly used, but they were later replaced or complemented with statistical methods

DOI: 10.1002/sdr

(Feder et al., 2022, Girju & Moldovan, 2002; Ittoo & Bouma, 2011). Whereas rule-based methods can achieve a high performance in the detection of explicit causality, machine learning-based methods surpass them in detecting implicit causality (Asghar, 2016).

As the name suggests, rule-based methods use a set of specific rules that enable them to extract causal information from text. For example, Khoo *et al.* (1998) defined a series of linguistic patterns that express causality using a list of causal markers. These markers consisted of causal verbs (*cause*, *break*, *kill* …), resultative constructions (such as (8) above), causal adverbs (*so*, *because*) and conditionals (see (6) and (7) above). Examples (11) and (12) illustrate patterns derived from these lists. The pattern in (11) uses the causal adverb *because*, whereas (12) contains a conditional construction marked by *if*:

11. *because [cause]*, *[effect]*
12. *if [cause]*, *[effect]*

A rule-based model can use these patterns to detect causality. For example, (13) is a causal sentence that can be found using the pattern in (11); (14) can be found using (12):

13. Because the chairs had broken, we had to eat standing up.
14. If the vase breaks, dad will be upset.

There are sundry other rule-based methodologies for detecting causality in the literature (e.g., Girju & Moldovan, 2002; Ittoo & Bouma, 2011). Among these, approaches like that of Khoo et al. (1998) depend on manually constructing linguistic patterns, while others employ automated methods for pattern generation. What all these methods have in common is that they rely on rules of the type *if x is found in a sentence*, *then the sentence is causal*. As a result, these systems have the capacity to detect explicit causality of a certain form but often not implicit causality. Implicit causality requires a reader to infer that the author meant to convey a causal relation between two concepts (as is the case for (9) and (10) above). Rules using the types of patterns given in (11) and (12) are incapable of detecting such expressions of causality.

Statistical methods diverge from rule-based approaches by not depending on explicit rules but rather employing data-driven techniques to create general-purpose or task-specific models. These methods assign vectors, simply put as arrays of numbers, to words based on their context within the text. This encoding, popularised by Mikolov et al. (2013), named word embeddings, enables a multitude of NLP applications (Liu et al., 2020). The emergence of the so-called pre-trained transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and Generative Pretrained Transformer (GPT) (Brown et al., 2020), has pushed state-of-the-art performances on most NLP tasks. Transformers were introduced by Vaswani et al. (2017) and enable the parallel processing of larger data sequences than previous architectures. They incorporate a self-attention mechanism that dynamically adjusts word representations based on their contextual relationship to other words, thereby improving their ability to capture long-range dependencies and context-specific

DOI: 10.1002/sdr

semantic relations. Consequently, predictions about word combinations are made with greater accuracy and efficiency.

BERT, for instance, considers associations in both forward and backward directions, facilitating better contextual understanding and identification of causal structures in sentences. Its architecture also supports task optimization, such as identifying causal sentences. BERT's adaptability for various tasks, including question answering and sentiment analysis, extends to causality detection through a process called fine-tuning. By exposing BERT to annotated data relevant to a specific task, it can deduce causal relationships even in previously unseen sentences. Recently, Tan et al. (2022) demonstrated the effectiveness of BERT for causality detection, leveraging its extensive training on diverse textual data sources.

Building on the work of Devlin et al. (2018) and Wolf et al. (2020), Tan et al. (2022) fine-tuned a BERT model for detecting causality with 3559 annotated sentences on causal events from news articles. The data included cases of explicit causality such as (15) as well as cases of implicit causality such as (16):

15. The bombing created panic among the villagers.
16. Dissatisfied with the package, workers staged an all-night sit-in.

Their fine-tuned BERT model was able to detect both explicit and implicit causality on the news corpora training data.

In this section, we have discussed rule-based and machine learning-based methods such as the BERT model, to extract causal sentences. The following section will introduce an experiment to illustrate and test the use of five NLP models to extract causal information to support model development.

METHOD

In this section, we describe a pipeline[2] and set of five NLP models for identifying causal information in text. We then describe three test cases and the process of collecting data from these cases. Finally, we describe the evaluation metrics used to assess the performance of the pipeline and models.

Architecture

Extraction of causal information from written data requires several steps. These steps have been combined in a pipeline that processes documents and outputs sentences labelled as either causal or not causal (Figure 1).

The input to the pipeline consists of a text document, which undergoes preprocessing to facilitate more efficient processing by NLP models. This involves 'cleaning' the text by removing unnecessary whitespace and other artefacts, followed by separating the text into individual sentences through a process called sentence tokenization. These sentences serve as the primary unit of analysis for BERT and GPT models. For the baseline method additional steps are required, as

[2]The code repository is available upon request.

DOI: 10.1002/sdr

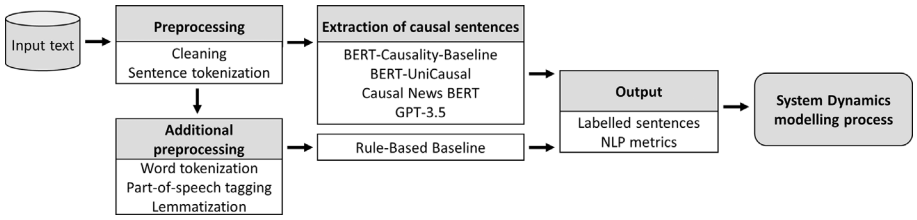

FIGURE 1. Pipeline architecture for causal information extraction. The model completeness metric was calculated outside of the pipeline.

this method does not consider the context of words in a sentence but only the presence of a word. For this reason, sentences are further tokenised into smaller units, known as tokens—that is, words, numbers, punctuation marks and other meaningful elements in the text. Each word is then processed with part-of-speech tagging, which assigns a classification such as noun, verb or adjective to each word. This facilitates the next step in the process: lemmatisation. Lemmatisation involves reducing different word forms to their base form. For instance, the forms 'understand', 'understands', 'understood' and 'understanding' are all forms of the same verb 'understand'. A lemmatiser detects these different forms and transforms them into the appropriate lemmatised stems (stripping inflexion and conjugation). This enables the baseline method to handle variants of a single word. For tokenisation, part-of-speech tagging and lemmatisation the corresponding functionality in the Python Natural Language Toolkit (NLTK) is used (Bird *et al.*, 2009).[3] The advanced methods require less preprocessing as they can utilise more information. Most importantly, they consider individual words in their complete form within the context of a sentence.

The subsequent step is causal information extraction; this process entails determining whether a given sentence has a causal meaning or not. For example, we want sentence (17) labelled '1' (causal) and sentence (18) labelled '0' (non-causal) in this step:

17. Viruses can cause a program to stop functioning. → 1
18. We all went to work that day. → 0

The model performance is evaluated using several metric scores implemented in the Scikit-learn package (F1, accuracy, precision, recall), which are discussed later (Pedregosa et al., 2011).[4]

Algorithms for causal relation identification

Currently, transformer-based pre-trained language models represent the state-of-the-art in NLP (Vaswani *et al.*, 2017). We compare three different methods for causality extraction, including a rule-based 'baseline' method and four transformer-based pre-trained language models: three BERT models fine-tuned for causal detection and GPT-3.5.

[3]Tokenisation: https://www.nltk.org/api/nltk.tokenize.html; part-of speech: https://www.nltk.org/api/nltk.stem.wordnet.html; lemmatisation: https://www.nltk.org/api/nltk.stem.wordnet.html.
[4]https://scikit-learn.org/stable/modules/model_evaluation.html.

DOI: 10.1002/sdr

The first model is a simple, rule-based baseline. The Baseline model uses a list of 215 causal verbs such as *cause*, *reduce* and *make* and verb phrases such as *have an effect*, *be due to* and *be responsible for*. The list was developed in three steps: first, we manually extracted causal verbs and verb phrases from a selection of documents. Subsequently, we enriched the list by finding similar words using Word2Vec: a model that was trained on approximately five million PubMed abstracts (Mikolov et al., 2013). The list was extended further with additional sets of causal verbs from Girju (2003) and Ittoo and Bouma (2011). The resulting list includes causal verbs typical for SD, such as *accumulate* and *influence*. This baseline model simply checks whether one of the verbs or verb phrases from the list is present in a sentence and assigns a causal or non-causal label.

Given that the list with causal verbs includes verbs from multiple sources and is enriched with similar words, we predict that this method is effective at identifying explicit causal sentences. The downside is that this approach can lead to inaccuracies. For instance, *make* was listed as one of the causal verbs, but leads to the non-causal sentence in (19) being labelled as causal. Furthermore, all cases of implicit causality will be missed since they are not indicated with a specific verb:

19. The leaflet contained information on the car's make and model.

For these reasons, we incorporated four more advanced NLP models. The first is the Causal News BERT model we described above (Tan et al., 2022). In addition, we test the BERT–causality–baseline, a BERT model trained by Nik et al. (2022) using similar training data. Furthermore, we use the UniCausal model (Tan et al., 2023). UniCausal is a fine-tuned BERT Sequence Classification model, which is a version of BERT that is specifically aimed at the task of determining if a text contains causal relations. The training and testing procedure of UniCausal is similar to that of the Causal News BERT model: the difference lies in the use of a combination of five training and test datasets.

Like BERT, the GPT model is based on a transformer architecture and pre-trained on a massive collection of texts (Brown et al., 2020). It is most known for its implementation in ChatGPT. We used the GPT-3.5 text-davinci-003 application programming interface.[5] An important difference between the BERT models used in this paper and the GPT model is that the latter has not been fine-tuned for causal identification. Rather we give GPT the prompt: Does the following sentence contain a cause–effect relation?[6]

Data collection

We selected three texts to illustrate the potential of causal extraction to inform the SD modelling process. In the following sections we will explore the efficacy of various NLP models to do so based on these texts. Our selection was

[5]https://platform.openai.com/docs/guides/gpt/completions-api.
[6]We have tested several other prompts: Does the sentence express causality? Does the following sentence express influence between two or more concepts? Is this sentence relevant for building a system dynamics model?

DOI: 10.1002/sdr

guided by four criteria: the presence of a causal theory, manageable length, method variety and topic variety. To ensure a sufficient number of relevant sentences to assess the NLP models' performance, we selected texts that discuss a causal theory. The need for manual annotation necessitated limiting both the number of cases and length of the texts to a degree which was manageable in the time available. We selected texts that differ in how they analyse and discuss causality, as this might impact the way causality is described (causal loop diagram, stock and flow model, political and historic analyses). Furthermore, language is context specific; therefore, three texts were selected that address different topics (transportation, climate change, foreign policy). The three selected cases are:

- Traffic congestion (TC) case: An excerpt from an educational book by Sterman (2000), which describes the problem of traffic congestion and demise of public transport using a causal loop diagram.
- Climate change (CC) case: An excerpt from a paper by Sterman and Booth Sweeney (2002), which describes the dynamics of the carbon cycle and global warming using a stock and flow model.
- Foreign policy (FP) case: A full paper by McFaul (2020) published in a foreign policy journal, which describes a theory on the determinants of Russian foreign policy.

The TC and CC texts provide a clear causal hypothesis made explicit in SD models. This enables us to assess how well the output of the NLP models captures the relations described in the texts and check the results against the SD model. The TC case discusses general causal relations, while the CC explicitly discusses stock-and-flow structures. Additionally, we have included the FP case from outside the field of SD. The FP case does not discuss an explicit SD model but does propose a causal theory using language that might be more common in other fields of science and policy documents. Thus, it serves as an example of the type of text that an SD modeller might use as source material. The three texts are preprocessed as described above.

Table 1 provides descriptive statistics on the content of each text. Each sentence is manually annotated by three of the authors. Labels are assigned according to our annotation guidelines. Differences between the annotations were observed, which is not uncommon in causal annotation. The differences were

TABLE 1. Test case data descriptive statistics.

	Sentences		
Cases	Total	Causal	Model relation
Traffic Congestion	151	71 (47%)	60 (40%)
Climate Change	137	57 (42%)	29 (21%)
Foreign Policy	549	199 (36%)	NA

Total sentences in case, number of sentences assessed as causal by annotation, number of sentences containing information on model relations (which are not all causal as we will discuss later).

DOI: 10.1002/sdr

discussed and a final label was assigned based on consensus. We base our annotation guidelines on the work of Dunietz *et al.* (2017), who build upon the work of Grivaz (2010):

1. **Why**: A "Why" question regarding the Effect can be constructed.
2. **Temporal order**: The Cause precedes the Effect in time.
3. **Counterfactual**: The Effect is not equally likely to occur or not occur without the Cause.
4. **Ontological asymmetry**: The Cause-and-Effect claims cannot be easily reversed.
5. **Linguistic**: A sentence is likely causal if it can be rephrased as "X causes Y" or "Due to X, Y."

Additionally, we added the following guidelines:

6. **Implicit Causality**: is annotated as causal: *[CAUSE] As population density falls*, *[/CAUSE]*, *[EFFECT] fewer and fewer people live near a bus or sub- way route [/EFFECT]*.
7. **WH questions (What, who, etc.) + yes/no-questions**: are annotated as causal: *[CAUSE] What [/CAUSE] determines [EFFECT] travel time [/EFFECT]?*
8. **Negative Causality**: is annotated as causal: *[CAUSE] Without such understanding [/CAUSE] [EFFECT] people are likely to rely on the intuitive 'wait and see' strategy that works well in a range of everyday tasks [/EFFECT]*.
9. **Embedded Causality**: is annotated as causal: *Small wonder that the challenge of implementing [CAUSE] policies [/CAUSE] [EFFECT] to reduce the impact of warming [/EFFECT] remains unanswered*.
10. **Inter-sentential causality** is NOT annotated as causal.

NLP evaluation metrics

To evaluate performance, we compare the NLP models' output with the manual annotation (causal: yes/no) and the extent to which model relations in the TC and CC are identified.

Four common metrics from the NLP field are used:

Accuracy expresses how often the model made a correct assessment:

$$\text{Accuracy} = \frac{\#\text{True positives} + \#\text{True negatives}}{\#\text{Total}}$$

Precision expresses to what extent sentences identified as causal by the model were indeed causal:

$$\text{Precision} = \frac{\#\text{True positives}}{\#\text{True positives} + \#\text{False positives}}$$

DOI: 10.1002/sdr

Recall expresses the extent to which the model identifies all causal sentences:

$$\text{Recall} = \frac{\#\text{True positive}}{\#\text{TruePositive} + \#\text{False negative}}$$

The F1 score provides a general measure of performance by equally combining precision and recall:

$$2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

In discussing our results, we tend to favour recall slightly because it helps us retrieve as much causal information as possible from a text. Overlooking such information might prevent us from identifying important relations for an SD model.

A causal relation relevant for an SD modeller might be mentioned in more than one sentence, while one sentence might contain information on more than one causal relation. This means that all relevant causal information could potentially be retrieved by an NLP model even if it fails to identify all causal sentences in a text. The TC and CC cases present an opportunity to assess this, as these texts include an explicit model that can be used as a 'ground truth' for relevant causal information, which is a subset of all causal information in the text. We can determine to what extent the NLP models find at least one sentence related to each relation included in the model. Stated differently: Had the model not been included in the case, could we have constructed it based on the NLP model output? To make this assessment we have annotated which model relations are referred to in each sentence. To assess the performance of the model we determine a 'completeness' score:

$$\text{Completeness} = \frac{\#\text{Model relations extracted atleast once}}{\#\text{Model relations mentioned in text}}$$

The completeness score describes whether the model has identified at least one sentence that contains information about each model relationship.

RESULTS

In this section, we review the ability of five NLP models to retrieve relevant information for SD modelling from three text cases. As shown in Table 1, a fairly large portion of the sentences in each text can be considered causal. The NLP models retrieve a share of this information, and the NLP metrics discussed below describe the extent to which they do. We discuss the results per case and reflect on the differences between the models and differences between the cases.

This also raises the question of whether all causal sentences in a text are relevant for model building. From Table 1, we can infer that a large share of the sentences in the texts is considered causal (TC: 47%; CC: 42%; FP: 36%).

DOI: 10.1002/sdr

Furthermore, the TC and CC case contain multiple causal sentences (TC: 21; CC: 35) that are not related to a model relation included in the case. This is an indication that not all causal information in a text is relevant for SD modellers. This leads us to ask if the NLP models retrieve the sentences that are relevant. We do this by reviewing the model completeness scores for the TC and CC texts. Finally, we investigate sentences that are not causal in nature but that do contain information about system structure.

NLP metrics

Tables 2–4 display the performance of the models in identifying sentences containing causal information in the TC, CC and FP cases. The TC case is relatively simple, with many causal relations being very explicitly mentioned in the text. Both the BERT–causality–baseline and Causal News BERT models perform similarly well in the TC case by achieving a 0.85 and 0.87 recall which indicates that a large share of the causal information in the text was detected (see Table 2). The models achieve this performance with a fair degree of precision (0.76 and 0.72). GPT-3.5 shows the poorest performance with a rather low recall of 0.41, which indicates that a large share of the causal information in the text was not identified.

TABLE 2. Performance on the Traffic Congestion (TC) case.

	F1	Accuracy	Precision	Recall
Rule-based Baseline	0.67	0.68	0.65	0.69
BERT–Causality–Baseline	0.80	0.80	0.76	0.85
BERT–UniCausal	0.71	0.74	0.76	0.66
Causal News BERT	0.79	0.78	0.72	0.87
GPT-3.5	0.55	0.69	0.85	0.41

In the CC case (Table 3) we again see the Causality–Baseline and Causal News BERT models performing well, with a recall of 0.75 and 0.72. In this case we also see the BERT–UniCausal model performing well, with a recall of 0.74 paired with the highest precision (0.82). Performance for most models is slightly lower in the CC compared to the TC case, except for GPT-3.5, which manages to achieve a slightly higher recall (0.53).

TABLE 3. Performance on the climate change (CC) case.

	F1	Accuracy	Precision	Recall
Rule-based Baseline	0.61	0.66	0.58	0.63
BERT–Causality–Baseline	0.75	0.80	0.79	0.72
BERT–UniCausal	0.78	0.82	0.82	0.74
Causal News BERT	0.74	0.77	0.72	0.75
GPT-3.5	0.63	0.75	0.79	0.53

DOI: 10.1002/sdr

All NLP models show their lowest precision and recall performance in the FP case, although the relative performance differences to other cases are mostly small (see Table 4). Since the TC and CC cases are SD texts, the underlying reason for this is likely that SD scholars more explicitly express causality, making it easier for the NLP models to label sentences correctly. Causal News BERT has the highest recall (0.77) combined with the second highest precision (0.68). BERT–Causality–Baseline also achieves fair performance, with a recall of 0.68 and a precision of 0.67.

TABLE 4. Performance on the foreign policy (FP) case.

	F1	Accuracy	Precision	Recall
Rule-based Baseline	0.53	0.62	0.48	0.60
BERT–Causality–Baseline	0.68	0.76	0.67	0.68
BERT–UniCausal	0.59	0.75	0.74	0.49
Causal News BERT	0.72	0.79	0.68	0.77
GPT-3.5	0.48	0.72	0.71	0.37

The performance differences between the two best-performing models, BERT–Causality–Baseline and the Causal News BERT models, are small. When purely focused on recall, Causal News BERT has the highest performance in three out of three cases. However, its precision is slightly lower compared to other advanced models, especially in the TC and CC. Based on a more balanced performance indicator, BERT–Causality–Baseline shows solid performance, with the highest F1 score in the TC, and second highest F1 score in the CC and FP. The BERT–UniCausal model suffers from mediocre recall performance in the TC and FP cases but otherwise performs well. The GPT-3.5 model struggles to identify causal sentences, which leads to the poorest recall performance in all cases; however, GPT-3.5 does have slightly higher precision than BERT–causality–baseline and Causal News BERT. The Baseline model has the poorest performance in accuracy and precision across the cases and has a significantly lower recall then the top-performing BERT model in each case. This illustrates the superior performance of advanced NLP techniques, especially fine-tuned models, compared to simple rule-based methods.

Model completeness score

A large variation exists in the number of times a relationship is referenced in the texts. In the TC case, a relationship is mentioned 4.2 times on average (min:1; max: 13). On average, a sentence that contains information about model relations includes details on 2.2 relations (min: 1; max: 6). The model completeness results are encouraging. Causal News BERT (100%) and BERT–Causality–Baseline (97%) are most effective, followed by GPT-3.5 (91%), BERT–UniCausal (91%) and Baseline (88%). Causal News BERT effectively identifies at least one relevant sentence for each model relationship mentioned in the text, and for many relationships it identifies all associated sentences (see Figure 2).

DOI: 10.1002/sdr

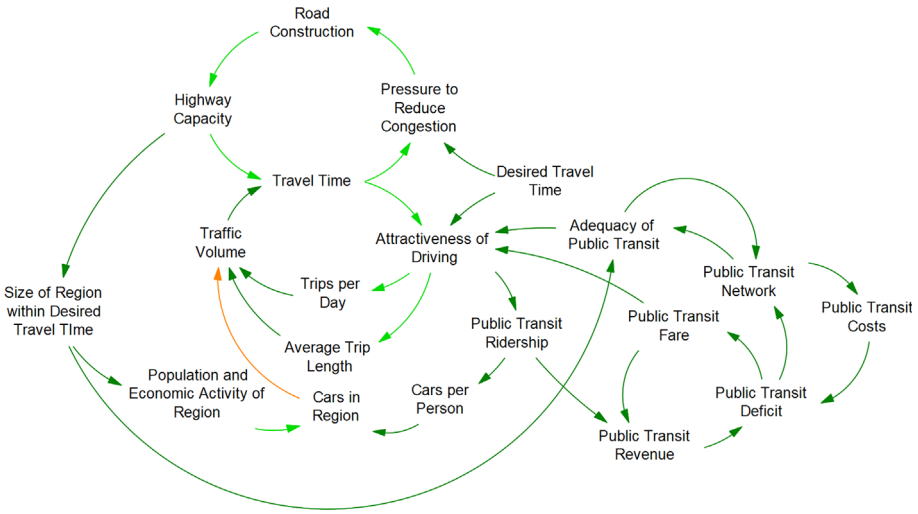

FIGURE 2. The TC case reference model. Colours indicate the percentage of sentences referencing the relationships that were extracted by the Causal News BERT model. Orange: 33%; light green: 66–92%; dark green: 100%. This is a modified version of Sterman (2000, figs 5–37, p. 187).

Identifying sentences that refer to relations in the CC model is more challenging; performance varies between 56% and 94%. Causal News BERT (94%) and BERT–Causality–Baseline (88%) are most successful in at least identifying one sentence per relationship in the model and perform noticeably better than the other model (GPT-3.5: 75%; BERT–UniCausal: 63%; Baseline: 56%). Sentences that include information about relations on average mention 1.9 relations (min: 1; max: 4). On average, each relationship is mentioned 2.9 times (min: 0; max: 10),[7] with four relationships only mentioned once. The best-performing models generally extract a large percentage of the sentences which reference a specific relationship (see Figure 3). The lower performance in the CC case compared to the TC case can be attributed to individual relationships being mentioned far less frequently in the CC case. This means that if an NLP model fails to identify certain sentences the SD modeller might overlook relations mentioned in the text. This presents a paradox in the results. BERT–UniCausal was relatively effective at finding model relations in the TC (87%) but performed relatively poorly in the CC (63%), despite having the second highest recall (0.74). BERT–UniCausal fails to identify several sentences that are crucial to identify a relation and overlooks relations that are mentioned in up to four different sentences.

Qualitative analysis: Causality and information about system structure

The CC case explicitly discusses stock and flow structure. When a flow is merely described as a transfer, such as '… the transfer of heat from the surface layer to the deep ocean' and '… the carbon in biomass is transferred to soils' (Sterman & Booth Sweeney, 2002, p. 212) they are typically not recognised as causal relations in NLP. Our annotation follows NLP guidelines and reflects this (what causes heat or carbon to flow?). Without a doubt, the two examples above contain relevant information for SD modellers about the structure of the system. NLP models

[7]Three relationships related to the variable 'Temperature Difference' are not mentioned in the text.

DOI: 10.1002/sdr

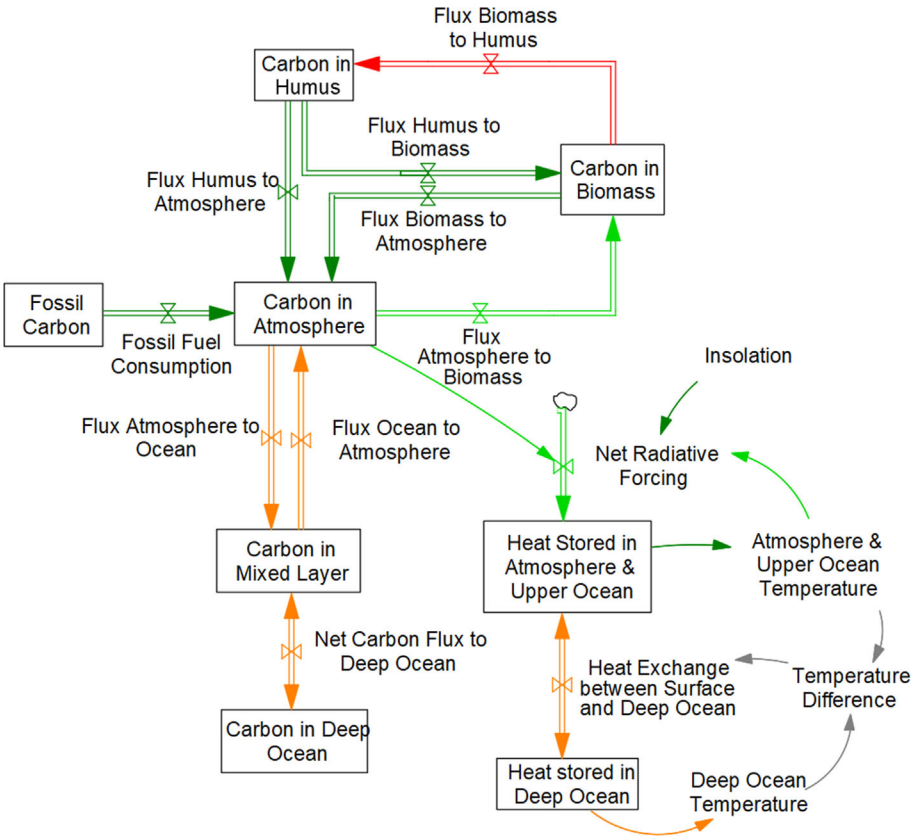

FIGURE 3. The CC case reference model. Colours indicate the percentage of sentences referencing the relationships that were extracted by the Causal News BERT model. Red: 0%; orange: 50%; light green: 75–86%; dark green: 100%; grey: not mentioned in the text. This is a modified version of Sterman and Booth Sweeney (2002, fig. 2, p. 213).

trained to identify causal sentences can thus fail to identify relevant information for SD models. For example, in the CC case, the latter example was the only sentence that contained information about the flow 'Flux Biomass to Humus'. This does not mean that NLP models overlook all information on flows, as most flows are discussed in relation to a causal effect. For example, all models identify the following sentence correctly as causal: 'An injection of fossil carbon to the atmosphere leads to a rise in average surface temperatures' (p. 212).

Other examples can be found of sentences that describe system structure but are not causal following the criteria used in NLP. For example, the TC contains several sentences about equations, such as: 'Total traffic volume must therefore equal the number of vehicles in the region multiplied by the number of miles each vehicle travels per day' (Sterman, 2000, p. 181). Although this equation contains information about the system's structure, it does not describe a specific causal relation between events nor behaviour-over-time. A third example of information about system structure which is not causal by criteria used in NLP can be found in the FP case: '… 470 IRA-controlled accounts responsible for 80,000 posts …' (McFaul, 2020, p. 135). This specific sentence contains important information on an actor and its activity but is unclear about the cause of the activity. However, again most sentences that describe actors and action are causal according to the NLP criteria used; for instance: 'The more consolidated Russian

DOI: 10.1002/sdr

FIGURE 4. An example of a small model structure inferred from a selection of sentences extracted from the FP case.

autocratic institutions became, the more influence Putin wielded individually on foreign policy' (p. 115).

Practical use of the output

The pipeline outputs a labelled list of sentences that users can import into a spreadsheet or other software application. This enables users to quickly filter and search the list for sentences that contain causal information. This information can then be used in the model-building process, for instance, by inferring a piece of model structure, as we will illustrate below.

The results demonstrate that the best-performing NLP models are effective at identifying a large portion of the causal sentences in three texts relevant for SD modellers. Causal News BERT is a top-performing model and reaches a model completeness score of 100% in the TC case and 94% in the CC case. Figures 2 and 3 illustrate the extent to which the NLP model has labelled the available intra-sentential information about specific relations as causal. Users aiming to construct models based on sentences labelled as causal would find comprehensive, often complete, descriptions of the model relations as far as available in the text.

However, a closer inspection of the results reveals that NLP models trained to retrieve causal information might, understandably, overlook some valuable information for describing system structure. This includes certain descriptions of flows, relations between actors and their actions and equations. Opportunities for mitigating this problem are described in the 'Discussion' section, below.

Although the FP case lacks a reference model, we can demonstrate how key variables and relations can be extracted from the text by constructing a small model that explains the change in Russian foreign interventions (see Figure 4). We do so by reviewing a number of causal sentences extracted by the Causal News BERT model.

In the FP case, McFaul argues that (2020, p. 100): 'Putin selected a unique trajectory for Russian foreign policy because of a set of particular ideas that he developed ….' While Russia's growing capability for foreign interventions was a factor of importance, 'New Russian capabilities did not make these Russian interventions inevitable' (p. 100). Rather, the definition of Russian national interests according to a set of ideas championed by President Putin are a key component (pp. 98–99): 'He embraced and propagated illiberal, conservative nationalism to

DOI: 10.1002/sdr

advance his definition of national interests.' However, President Putin's growing autonomy and control over Russian foreign policy were essential for acting on these ideas, adding three more causes to the argumentation (p. 100): 'The Russian system of government became increasingly autocratic during Putin's two decades of rule, giving Putin more autonomy and more influence over Russian foreign policy.' From McFaul's analysis, we can construct a small model that summarises some of the key causal relations described in the paper (see Figure 4). This small model could serve as the starting point for a model or as an addition to an existing model. While the range and depth of causal mechanisms McFaul discusses goes beyond this brief example, it illustrates how a piece of model structure can be derived from extracted sentences. The FP case example also illustrates that a gap still exists between the output of the NLP models tested in this study and a SD model. Currently, it is the modeller who has to identify relevant variables and relations in the sentences labelled as causal. We will revisit this gap in the 'Discussion' section, below. Nonetheless, the three case studies provide preliminary confidence that NLP models can be used to support model development by retrieving relevant information from texts.

DISCUSSION

In this paper, we explore how NLP models can be used to extract sentences containing causal information from texts, marking a step towards incorporating artificial intelligence into the SD toolkit. We provide a primer on the notion of causality in SD and demonstrate how causal information can be identified in text, a topic that has received limited attention in SD literature. We find that NLP models are effective at identifying a large portion of the causal sentences in texts. More advanced models, such as fine-tuned BERT models, outperform a simple baseline model. Specifically, the Causal News BERT model was effective at reaching reasonable levels of accuracy (TC: 0.78; CC: 0.77; FP: 0.79), precision (TC: 0.72; CC: 0.72; FP: 0.68) and recall (TC 0.87; CC: 0.75; FP: 0.77) across cases, while also identifying 100% and 94% of the model relations contained in the TC and CC cases. The differences with the BERT–Causality–Baseline model are small. Constant advancements in NLP technology and the potential for task-specific fine-tuning means that the performance of future models is likely to exceed the results presented here. A closer inspection of the results reveals that NLP models trained to extract causal information overlook non-causal information which is relevant for describing system structure, such as the description of flows, activities and equations. Nonetheless, this study provides preliminary confidence that NLP models can retrieve information to support model development. However, several challenges remain. Below, we will discuss the challenges and potential solutions for retrieving a broader set of relevant information from text and enhancing the practical value of NLP tooling. We also discuss the importance of human guidance and interpretation during model development.

Understanding the operations of a system is fundamental to the SD method. However, we observe that two key elements are not inherently causal based on criteria used in NLP: activities and flows. Richmond's (1993) and Olaya's (2016) now somewhat famous example describes that models of milk production should

DOI: 10.1002/sdr

include cows. However, the sentence: 'cows produce milk' is not causal, although it has an instrumental property embedded in it. Citing Machamer *et al.* (2000), Olaya (2016) states: 'An entity acts as a cause when it engages in a productive activity' (p. 202). Adhering to a slightly stricter definition of causality, common in NLP, the sentence lacks a clear cause in the form of an event or behaviour-over-time explaining why cows produce milk. Concerning flows, the text fragment 'transfer of heat from the surface layer to the deep ocean' provides valuable information about an important flow in the system. But again, it lacks clear causation and as a result we and NLP models consider the sentence to be non-causal. These limitations barely affect the overall performance in the current analysis since most descriptions of an activity or flow do include a cause. Nonetheless, descriptions of activities and flows are clearly relevant pieces of information for SD modellers. This leads us to propose several avenues for extending the current approach.

While our initial approach focused on causal information, broader capabilities are both available and necessary. The pipeline can be improved by using improved NLP models or by adding additional task-specific NLP models. NLP models can be designed and fine-tuned to specifically identify information relevant for SD modelling, including feedback loops, nonlinear relationships, flow dynamics, stock variables and the operations that interconnect them. Furthermore, the analysis can be extended to extract actor behaviour, such as the following sentence from the FP case: 'Over time, however, Putin grew more suspicious of private economic actors …' (McFaul, 2020, p. 110). This can be accomplished by using existing training data and models as well as a joint effort by SD researchers to curate and annotate a training set to fine-tune NLP models for SD applications. This is a realistic venture, considering that the causal news corpus used to train two of the BERT models used here consists of 'only' 3559 events (Tan et al., 2022). The NLP field offers starting points for such an endeavour. Bakker et al. (2022a, 2022b) and Sil *et al.* (2010) have developed methods to extract combinations of actions, actors, objects, recipients, preconditions and postconditions from texts. The output of these methods can benefit operational thinking during SD modelling by identifying actors, their decisions, and subsequent actions and results, alongside the information relations that guide these processes. Additionally, Talmy (1988) described the semantics of force dynamics, including flows, which could provide a starting point to operationalise how flows and other dynamics can be identified in text. In addition, to further enhance information extraction, specific models can be included in the pipeline that focus on inter-sentential causality (Jin *et al.*, 2020).

The current pipeline and NLP models provide an initial method for modellers to extract a significant portion of sentences containing causal information. Yet, the analysis still heavily depends on the work by the modeller since the NLP models lack the capability to synthesise information across text segments or to identify variables and relationships for direct use in modelling. Various qualitative research techniques for model development are available to assist a modeller in this task, but great potential lies in further automatization (see Eker & Zimmermann, 2016; Kim & Andersen, 2012; Luna-Reyes & Andersen, 2003; Tomoaia-Cotisel et al., 2022; Turner et al., 2013; Yearworth & White, 2013). Selecting appropriate variables and relations to include in a model touches upon

DOI: 10.1002/sdr

the essence of model building and requires a nuanced understanding of the model's purpose and problem context. Acquiring and articulating such understanding can be challenging for the modeller, especially in the early stages of model development. Thus, transferring this understanding to an NLP model to guide variable and relation selection and aggregation is problematic. Furthermore, NLP models that are to some extent capable of interpreting the task they have been assigned suffer from deficiencies such as hallucination, bias and distraction (Shi et al., 2023; Zhao et al., 2023). We therefore believe that, for now, the path forward lies in several developments that can advance our initial approach into a valuable and practical support tool, under the supervision of the modeller. We will explore these developments in the context of causal relations, but they may also be applied to other types of semantic relations, such as flows or interactions between actors.

To facilitate merging results from multiple text fragments into a more extensive map of dependencies, the cause and effect must first be isolated from a text fragment; this process is known as span detection. This task is considerably more challenging than detecting causality (Mostafazadeh et al., 2016; Tan et al., 2023). Specific models for this task are available, like an adaptation of the BERT–UniCausal model. However, the cause span ('I hire transportation') and effect span ('and my customers have fresh cold cuts every day') detected by BERT–UniCausal (Tan et al., 2023, p. 2) are also still far removed from variables formulated as nouns or noun phrases, as is common in SD models (e.g., 'transport availability' and "fresh deliveries"). In addition, sources may use different vocabulary or discuss information at different levels of aggregation, complicating the integration of results from various text fragments. NLP techniques can identify words that are similar or belong to sub-classes (such as transportation, car, van), which can help in recognising causes and effects with similar meanings and overarching concepts (see Mikolov et al., 2013). However, it is both unlikely and undesirable that NLP models will output ready-to-use models directly—an intermediary step is required.

A promising development in NLP to tackle this challenge is the utilisation of knowledge graphs (Ji et al., 2021; Khadir et al., 2021). As described by Ji et al. (2021, p. 1): 'A knowledge graph is a structured representation of facts, consisting of entities, relationships, and semantic descriptions.' They visualise a layered ontology of concepts, synonyms and a variety of semantic relations between them, such as causal relations. Knowledge graphs can serve as an intermediary step between NLP models and the SD model-building process, offering an accessible way of combining, structuring and visualising results, while also maintaining traceability to source materials. Concepts and relations in a knowledge graph can be input from one NLP model or a collection of NLP models, each focused on extracting specific concepts or semantic relations. Hybrid forms of knowledge graph creation are being developed that allow users to evaluate and direct the graph construction process (Bakker et al., 2022a, 2023; de Boer & Verhoosel, 2020; Opasjumruskit et al., 2022). This could enable a modeller to guide the process based on the purpose of the SD model and other considerations—for instance, by indicating which concepts in the graph are relevant, irrelevant or can be aggregated. Extracting relevant concepts and semantic relations and combining them in a knowledge graph is akin to constructing a

DOI: 10.1002/sdr

proto-model. Integrating these advancements in a practical tool brings us close to the vision articulated by Sterman in 2000 (p. 898): 'a model-building tutor' capable of suggesting concepts and relations to include in a model.

Progress is possible, but we must proceed with some caution. Written data is an interpretation by the original authors and shaped by their access to information, motives, assumptions and other factors. In some cases, this subjectivity might be important for the modeller to include; in other cases it can lead to dangerous biases. Human interpretation will remain essential. As Forrester (1980, p. 557) explains:

> "To be useful, the literature must be pieced together, decisions must be interpreted into policies … One must read between the lines … It may be that such interpretation of the … literature cannot be effectively done without first-hand knowledge of the mental data base used by operators … Such first-hand knowledge can be obtained only by living and working where the decisions are made and by watching and talking with those who run the … system."

The use of artificial intelligence to support SD model development can represent a profound leap forward for the field. Thus far, the convergence between the fields of artificial intelligence and SD has been limited (Armenia et al., 2024). This article takes an initial step by focusing on the extraction of causal sentences from a text using NLP models. In the discussion, we have outlined several steps to refine NLP models for SD use, aiming to create practical tools that enable modellers to create proto-models. These tools can assist in processing extensive collections of texts, such as academic literature, company documents or government reports. Subsequently, they can be used to reconstruct, compare, integrate and analyse the system structure described in texts, aiding in the development of SD models. Presently, this process relies on the laborious reading and analysing of documents by SD modellers. The proposed advancement would enable modellers to thoroughly process more information in less time, ultimately leading to better models that are more capable of tackling the urgent issues addressed by the SD field.

ACKNOWLEDGEMENTS

We would like to thank Stephan Raaijmakers and Koen van der Zwet from TNO, and Etiënne Rouwette from Radboud University for their feedback on earlier drafts of the manuscript. We also thank the Editor and anonymous reviewer for their invaluable feedback during the writing process.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

DOI: 10.1002/sdr

Biographies

Guido Veldhuis is a Senior Scientist at the Defence, Safety and Security unit of TNO, The Netherlands. He holds an MSc from the Erasmus Mundus Programme in System Dynamics and is currently pursuing a PhD as an external candidate at Radboud University. His research focuses on developing methods and models to support decision-makers facing complex social and security-related challenges. Guido has served as the chair of the Benelux System Dynamics Society chapter.

Dominique Blok is a researcher who works on the development of large language models (LLMs) at TNO. She holds a PhD in linguistics from Utrecht University and she currently focuses on data curation methods to increase the representation of different groups in LLM datasets in order to increase diversity and reduce bias in LLMs.

Maaike de Boer (PhD) is a senior scientist at TNO within the Data Science Department. She obtained her BSc and MSc degrees in artificial intelligence (AI) (with a minor in linguistics) at the Utrecht University and her PhD at the Radboud University Nijmegen. At TNO, Maaike focuses on natural language processing and hybrid AI, combining data-driven methods and knowledge-driven methods to get the best of both worlds.

Gino Kalkman is a researcher at TNO working on the application of natural language processing (NLP) techniques in the biomedical domain. He obtained his PhD at the Vrije Universiteit Amsterdam in the field of biblical studies. At TNO, Gino's research focuses on the implementation and integration of NLP algorithms in scientific pipelines and user-friendly tools.

Roos Bakker (MSc) is a data scientist at TNO within the Data Science Department, and a PhD student at Leiden University. She obtained an MSc in artificial intelligence from Utrecht University and is currently pursuing a PhD in collaboration with TNO and Leiden University on the topic of knowledge graph extraction, enrichment, and evaluation. Roos is experienced in manual ontology development and natural language processing and is passionate about combining both to advance the field of knowledge representation and extraction.

Rob van Waas is a researcher at TNO specialising in complex systems modelling applied to defence and security issues such as migration, climate security and defence capability planning. In this role he enjoys the balance between research and application. He is a TU Delft systems engineering and policy analysis alumnus and gained experience in different positions at L'Oréal in the Benelux before switching to a research role at TNO.

DOI: 10.1002/sdr

REFERENCES

Akkasi A, Moens MF. 2021. Causal relationship extraction from biomedical text using deep neural models: a comprehensive survey. *Journal of Biomedical Informatics* **119**(103): 820.

Armenia S, Franco E, Iandolo F, Maielli G, Vito P. 2024. Zooming in and out the landscape: artificial intelligence and system dynamics in business and management. *Technological Forecasting and Social Change* **200**(123): 131.

Asghar N 2016. Automatic extraction of causal relations from natural language texts: a comprehensive survey. arXiv preprint arXiv:1605.07895.

Bakker RM, de Boer MH, Meyer-Vitali AP, Bakker BJ, Raaijmakers SA. 2022a. A hybrid approach for creating knowledge graphs: recognizing emerging technologies in Dutch companies. *HHAI2022: Augmenting Human Intellect* (pp. 307–309). OS Press: Amsterdam, The Netherlands.

Bakker R, van Drie RA, de Boer M, van Doesburg R, van Engers T. 2022b. Semantic role labelling for dutch law texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 448–457). European Language Resources Association: Marseille, France.

Bakker RM, Kalkman GJ, Tolios I, Blok D, Veldhuis GA, Raaijmakers S, de Boer MHT. 2023. Exploring knowledge extraction techniques for system dynamics modelling: Comparative analysis and considerations. In *Proceedings of the 2023 BNAIC/BeNeLearn Conference*. Delft, The Netherlands.

Bird S, Klein E, Loper E. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc: Sebastopol, CA.

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P *et al.* 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* **33**: 1877–1901.

de Boer M, Verhoosel JP. 2020. Towards data-driven ontologies: A filtering approach using keywords and natural language constructs. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 2285–2292). European Language Resources Association: Marseille, France.

Devlin J, Chang MW, Lee K, Toutanova K 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dunietz J, Levin L, Carbonell JG. 2017. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*. Valencia, Spain: Association for Computational Linguistics 95–104.

Eker S, Zimmermann N. 2016. Using textual data in system dynamics model conceptualization. *Systems* **4**(3): 28.

Feder A, Keith KA, Manzoor E, Pryzant R, Sridhar D, Wood-Doughty Z, Eisenstein J, Grimmer J, Reichart R, Roberts ME, Stewart BM. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics* **10**: 1138–1158.

Forrester JW. 1980. Information sources for modeling the national economy. *Journal of the American Statistical Association* **75**(371): 555–566.

Forrester JW. 1992. Policies, decisions and information sources for modeling. *European Journal of Operational Research* **59**(1): 42–63.

Girju R. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*. Sapporo: Association for Computational Linguistics; 76–83.

Girju R, Moldovan DI. 2002. Text mining for causal relations. In *Proceedings of the Fifteenth International Florida Artificial Intelligence ResearchSociety Conference*.

DOI: 10.1002/sdr

Pensacola Beach, Florida: Association for the Advancement of Artificial Intelligence (AAAI); 360–364.
Grivaz C. 2010. Human Judgements on Causation in French Texts. In *Proceedings of LREC 2010*. Valletta, Malta: European Language Resources Association (ELRA).
Gusenbauer M. 2019. Google scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics* **118**(1): 177–214.
Ittoo A, Bouma G. 2011. Extracting explicit and implicit causal relations from sparse, domain-specific texts. In *Natural Language Processing and Information Systems: 16th International Conference on Applications of Natural Language to Information Systems, NLDB 2011, Alicante, Spain, June 28–30, 2011. Proceedings 16*. Springer: Berlin Heidelberg; 52–63.
Ji S, Pan S, Cambria E, Marttinen P, Philip SY. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems* **33**(2): 494–514.
Jin X, Wang X, Luo X, Huang S, Gu S. 2020. Inter-sentence and implicit causality extraction from chinese corpus. In: Lauw, H., Wong, RW., Ntoulas, A., Lim, EP., Ng, SK., Pan, S. (eds.), *Advances in Knowledge Discovery and Data Mining: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (Part I)*. Cham: Springer Nature; 739–751. https://doi.org/10.1007/978-3-030-47426-3_57
Khadir AC, Aliane H, Guessoum A. 2021. Ontology learning: Grand tour and challenges. *Computer Science Review* **39**: 100,339 (2021).
Khoo CS, Kornfilt J, Oddy RN, Myaeng SH. 1998. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing* **13**(4): 177–186.
Kim H, Andersen DF. 2012. Building confidence in causal maps generated from purposive text data: Mapping transcripts of the Federal Reserve. *System Dynamics Review* **28**(4): 311–328.
Kunc M, Barnabè F, Giorgino MC. 2023. Uncovering dynamic complexity in annual reports: A methodological approach using resource mapping. *System Dynamics Review* **39**(4): 299–335.
Lane DC, Husemann E. 2010. What does the arrow mean? Observations on system dynamics mapping and the potential for experimentation with other methods. In: Strohhecker, J., Größler, A. (eds.), *Strategisches und operatives Produktionsmanagement*. Wiesbaden, Germany: Gabler Verlag.
Liu Q, Kusner MJ, Blunsom P 2020. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*.
Luna-Reyes LF, Andersen DL. 2003. Collecting and analyzing qualitative data for system dynamics: Methods and models. *System Dynamics Review* **19**(4): 271–296.
Machamer P, Darden L, Craver CF. 2000. Thinking about mechanisms. *Philosophy of science* **67**(1): 1–25.
McFaul M. 2020. Putin, Putinism, and the domestic determinants of Russian foreign policy. *International Security* **45**(2): 95–139.
Meadows DH. 2008. *Thinking in Systems: A Primer*. White River Junction, VT: chelsea green publishing.
Mikolov T, Chen K, Corrado G, Dean J 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
Mostafazadeh N, Grealish A, Chambers N, Allen J, Vanderwende L. 2016. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*. San Diego, CA: Association for Computational Linguistics; 51–61.
Nedjalkov P, Silnickij G. 1973. The topology of causative constructions. *Folia Linguistica* **6**: 273–290.

DOI: 10.1002/sdr

Nik A, Zhang G, Chen X, Li M, Fu J 2022. 1Cademy@ causal news corpus 2022: Leveraging self-training in causality classification of socio-political event data. *arXiv preprint arXiv: 2211.02729*.

Olaya C. 2016. Cows, agency, and the significance of operational thinking. *System Dynamics Review* **31**(4): 183–219.

Opasjumruskit K, Böning S, Schindler S, Peters D. 2022. Ontohuman: Ontology-based information extraction tools with human-in-the-loop interaction. In *Inter-National Conference on Cooperative Design, Visualization and Engineering*. Berlin: Springer-Verlag; 68–74.

Pedercini M. 2006. What's behind the blue arrow? The notion of causality in system dynamics. In *Proceedings of the 24th International Conference of the System Dynamics Society*. Nijmegen: System Dynamics Society.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O *et al.* 2011. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research* **12**: 2825–2830.

Pustejovsky J, Castano JM, Ingria R, Sauri R, Gaizauskas RJ, Setzer A *et al.* 2003. TimeML: Robust specification of event and temporal expressions in text. *New Directions in Question Answering* **3**: 28–34.

Richmond B. 1993. Systems thinking: Critical thinking skills for the 1990s and beyond. *System dynamics review* **9**(2): 113–133.

Schaffernicht M. 2010. Causal loop diagrams between structure and behaviour: A critical analysis of the relationship between polarity, behaviour and events. *Systems Research and Behavioral Science* **27**(6): 653–666.

Shi F, Chen X, Misra K, Scales N, Dohan D, Chi EH *et al.* 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*. Honolulu, HI: PMLR; 31210–31227.

Sil A, Huang F, Yates A. 2010. Extracting action and event semantics from web text. In *2010 AAAI Fall Symposium Series*. Arlington, Virginia: Association for the Advancement of Artificial Intelligence.

Sterman J. 2000. *Business Dynamics*. New York, NY: McGraw-Hill, Inc.

Sterman J. 2018. System dynamics at sixty: The path forward. *System Dynamics Review* **34**(1–2): 5–47.

Sterman JD, Booth Sweeney L. 2002. Cloudy skies: Assessing public understanding of global warming. *System Dynamics Review: The Journal of the System Dynamics Society* **18**(2): 207–240.

Talmy L. 1988. Force dynamics in language and cognition. *Cognitive science* **12**(1): 49–100.

Tan FA, Hürriyetoğlu A, Caselli T, Oostdijk N, Nomoto T, Hettiarachchi H, Ameer I, Uca O, Liza FF, Hu T 2022. The causal news corpus: Annotating causal relations in event sentences from news. *arXiv preprint arXiv:2204.11714*.

Tan FA, Zuo X, Ng SK 2023. UniCausal: Unified benchmark and model for causal text mining. *arXiv preprint arXiv:2208.09163*.

Tomoaia-Cotisel A, Allen SD, Kim H, Andersen D, Chalabi Z. 2022. Rigorously interpreted quotation analysis for evaluating causal loop diagrams in late-stage conceptualization. *System Dynamics Review* **38**(1): 41–80.

Turner BL, Kim H, Andersen DF. 2013. Improving coding procedures for purposive text data: Researchable questions for qualitative system dynamics modeling. *System Dynamics Review* **29**(4): 253–263.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN *et al.* 2017. Attention is all you need. *Advances in Neural Information Processing Systems* **30**: 5998–6008.

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A *et al.* 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on*

DOI: 10.1002/sdr

Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics; 38–45.

Yang J, Han SC, Poon J. 2022. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems* **64**(5): 1161–1186.

Yearworth M, White L. 2013. The uses of qualitative data in multimethodology: Developing causal loop diagrams during the coding process. *European Journal of Operational Research* **231**(1): 151–161.

Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y *et al*. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

DOI: 10.1002/sdr