5/10

2019_10pages_2021.10.30

- some references might be useful
- not very comprehensive / concrete / implementable ...

Identification of Causal Dependencies by using Natural Language Processing: A Survey

Erika Nazaruka[a]

Department of Applied Computer Science, Riga Technical University, Sētas iela 1, Riga, Latvia

Keywords: System Modelling, Knowledge Extraction, Natural Language Processing, Topological Functioning Model.

Abstract: Identification of cause-effect relations in the domain is crucial for construction of its correct model, and especially for the Topological Functioning Model (TFM). Key elements of the TFM are functional characteristics of the system and cause-effect relations between them. Natural Language Processing (NLP) can help in automatic processing of textual descriptions of functionality of the domain. The current research illustrates results of a survey of research papers on identification and extracting cause-effect relations from text using NLP and other techniques. The survey shows that expression of cause-effect relations in text can be very different. Sometimes the same language constructs indicate both causal and non-causal relations. Hybrid solutions that use machine learning, ontologies, linguistic and syntactic patterns as well as temporal reasoning show better results in extracting and filtering cause-effect pairs. Multi cause and multi effect domains still are not very well studied.

1 INTRODUCTION

Models are known from ancient times. Models are built for a specific purpose, and this determines their level of abstraction, accuracy, representation means, scale etc.

Traditional industries use graphical models for design and experiments to predict how a new product will function in the real circumstances. Software development models are mostly textual starting from requirements specifications and ending with the source code. Graphical models are used mostly to simplify understanding of key characteristics of the product.

The idea of using models as a core element of software development was met with interest when The Object Management Group had published their guide on Model Driven Architecture (MDA) in 2001 (Miller and Mukerji, 2001). MDA suggests using a chain of model transformations, namely, from a computation independent model (CIM, mostly textual) to a platform independent model (PIM, mostly graphical), then to a platform specific model (PSM, graphical) and to source code. The weaker place in this chain of transformations is the CIM, and its transformations. The CIM is dedicated for presentation of software requirements, business requirements, knowledge about the system domain, business rules, etc. The main task here is to process textual descriptions, graphical information, discover implicit knowledge, or, in other words, to find out all knowledge about system (software) functioning, behavior and structure. Analysis of the available information includes so called causal reasoning (Khoo et al., 2002). Identified causal dependencies are those of control flows in the systems and influence also some structural relations.

In our vision of implementation of MDA principles, we suggest using a knowledge model based on the Topological Functioning Model (TFM) as the CIM to generate code via an intermediary model – Topological UML model (Osis and Donins, 2017). The TFM elaborated by Janis Osis at Riga Technical University (Osis, 1969) specifies a functioning system from three viewpoints – functional, behavioural and structural. Causal dependencies are the key element in the Topological Functioning Model.

Construction of the TFM is based on analysis of verbal descriptions – instructions, interview protocols, position descriptions, or other experts' knowledge expressed in text. At the present we have

[a] https://orcid.org/0000-0002-1731-989X

Nazaruka, E.
Identification of Causal Dependencies by using Natural Language Processing: A Survey.
DOI: 10.5220/0007842706030613
In *Proceedings of the 14th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2019)*, pages 603-613
ISBN: 978-989-758-375-9

two approaches:

- manual processing of the text in the TFM4MDA (Topological Functioning Model for MDA) approach (Osis et al., 2007a; Osis and Asnina, 2011b) and
- automated processing of steps in use case scenarios in the IDM (Integrated Domain Modelling) toolset (Osis and Slihte, 2010; Slihte et al., 2011).

In practice, preparation of text and manual knowledge acquisition are too resource-consuming (Elstermann and Heuser, 2016). It is better either to skip the step of preparation of the textual description and start from human analysis of the available information, either to automate or semi-automate this process.

As metnioned, certain causal dependencies in a domain form control and message flows in a software system. This relates not only to the TFM (Nazaruka, 2017), but also to other models used in the transformations from CIM to (Kardoš and Drozdová, 2010; Kriouile et al., 2013; Kriouile et al., 2014; Kriouile et al., 2015; Bousetta et al., 2013; Rhazali et al., 2016; Essebaa and Chantit, 2016).

The key aspect of successful construction of the domain model is correct and complete identification of causes and effects. In software models, relations between causes and effects are implemented as control flows, message flows, transitions between states of the system. Not less important it is in case of the TFM construction, where identification of causal dependencies (topological relations) between functional characteristics of the system is crucial. The open question is how to identify and extract these relations from textual descriptions in the automated way that Natural Language Processing (NLP) tools suggest.

The goal of the given research is to make a survey on ways and completeness of extracting causal dependencies from text using NLP, Natural Language Understanding (NLU) and linguistics techniques.

The paper is organized as follows. Section 2 describes the purpose of the research and enumerates the research questions. Section 3 presents overview of research results on identification and extracting cause-effect relations from text. Section 4 presents a discussion on findings. Section 5 concludes the paper with discussion on main results and future research directions.

2 BACKGROUND AND RESEARCH QUESTIONS

This section discusses the background on cause-effect relations in the TFM, i.e. the brief overview of the TFM as well as formal definitions of cause-effect relations are presented. At the end of the section research questions are formulated.

2.1 Cause-effect Relations in the TFM

The TFM is a formal mathematical model that represents system functionality in a holistic manner. It describes the functional and structural aspects of the software system in the form of a directed graph (X, θ), where a set of vertices X depict functional characteristics of the system named in human understandable language, while θ is a set of edges that depict cause-effect relations (topology) between them. Such specification is more perceived, precise and clearer then the large textual descriptions. The TFM is characterized by the topological and functioning properties (Osis and Asnina, 2011a). The topological properties are connectedness, neighbourhood, closure, and continuous mapping. The functioning properties are cause-effect relations, cycle structure, inputs, and outputs. The composition of the TFM is presented in (Osis and Asnina, 2011b).

Rules of composition and derivation processes within TFM4MDA from the textual description of system functionality is provided by examples and described in detail in (Asnina, 2006; Osis et al., 2007b; Osis et al., 2008; Osis and Asnina, 2011b). The TFM can also be generated automatically from the business use case scenario specifications, which can be specified in the IDM toolset (Šlihte and Osis, 2014). It also can be manually created in the TFM Editor from the IDM toolset.

The cause-effect relations are those of causal dependencies that exist between functional characteristics of the system and define the cause from which the triggering of the effect occurs. In fact, this kind of relations indicates control flow transition in the system. For instance, termination of execution of a functional characteristic triggers initiation of execution of related characteristics (Figure 1).

Formal definitions of a cause-effect relation and a logical relation among those relations as well as their incoming and outgoing groups are as follows (Asnina and Ovchinnikova 2015; Osis and Donins 2017).

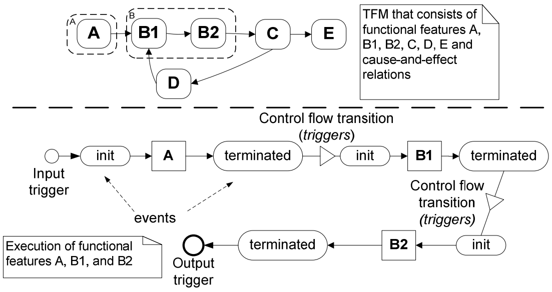

Figure 1: Execution of functional feature instances (Nazaruka et al., 2016).

Formal Definition of a Cause-Effect Relation. A cause-and-effect relation T_i is a binary relationship that relates exactly two functional features X_c and X_e. Both X_c and X_e may be the same functional feature in case of recursion. The synonym for cause-effect relation is a topological relationship. Each cause-effect relation is a unique 5-tuple (1):

$$T_i = \langle ID, X_c, X_{e,} N, S\rangle, \text{ where} \quad (1)$$

- ID is a unique identifier of the relation;
- X_c is a cause functional feature;
- X_e is an effect functional feature;
- N is a Boolean value of the necessity of X_c for generating X_e;
- S is a Boolean value of the sufficiency of X_c for generating X_e.

Formal Definition of a Logical Relation. A logical relation L_i specifies the logical operator *conjunction (AND), disjunction (OR),* or *exclusive disjunction (XOR)* between two or more cause-effect relations $T_{i.}$ The logical relation denotes system execution behaviour, e.g. decision making, parallel or sequential actions. Each logical relation is a unique 3-tuple (2):

$$L_i = \langle ID, \mathbf{T}, R_T\rangle, \text{ where} \quad (2)$$

- ID is a unique identifier of the relation;
- **T** is a set of cause-effect relations $\{T_i, ..., T_n\}$ that participate in this logical relation;
- R_T is a logical operator AND, OR, or XOR over **T**; operator OR is a default value.

Formal Definition of Incoming Topological Relations. A set of logical relations that join cause-and-effect relations, which go into functional feature X_i, is defined as a subset $\boldsymbol{L_{in}}$ of set $\boldsymbol{L} = \{L_i, ..., L_n\}$, where at least one topological relation T_i such that its effect functional feature X_e is equal to X_i is found in set T of topological relations in each logical relation L_i.

Formal Definition of Outgoing Topological Relations. A set of logical relations that join cause-and-effect relations, which go out from functional feature X_i, is defined as a subset $\boldsymbol{L_{out}}$ of set $\boldsymbol{L} = \{L_i, ..., L_n\}$, where at least one topological relation T_i such that its cause functional feature X_c is equal to X_i is found in set T of topological relations in each logical relation L_i.

The connection between a cause and an effect is represented by a certain conditional expression, the causal implication. It is characterized by the nature or business laws (or rules) not by logic rules. In causal connections "something is allowed to go wrong", whereas logical statements allow no exceptions. Using this property of cause-effect relations a logical sequence, wherein the execution of the precondition guarantees the execution of the action, can be prescinded, this means that even if a cause is executed, the corresponding effect can be not generated because of some functional damage.

The human mind applies very sophisticated mechanism as well as empirical information and world knowledge to construct "a theory of the causal mechanism that produced the effect" (Khoo *et al.* 2002). Trying to discover this "causal mechanism" they analyse "causal power" of events to generate an effect.

Since a cause generates an effect, a cause chronologically precedes an effect. This means that the cause-effect conditions contain a time dimension.

Causes can be sufficient or necessary (or complete or partial, correspondingly). A sufficient (complete) cause generates its effect ever, or in any conditions. On the other hand, a necessary cause (partial) only promotes its effect generating, and this effect is realized only if this partial cause joins other conditions. However, most cause-effect relations involve multiple factors. Sometimes there are factors in series. Sometimes there are factors in parallel. In case of the TFM, it is assumed that a deal is always with necessary causes as the functionality of the system has its known and unknown risks at the time of analysis.

A cause not only precedes an effect and always is followed by it, it causes (gives rise to, generates) and is condition on an effect. The concept of causing (generating) is necessary to distinguish a cause-and-effect relation from the simple consequence that is not causal. The causality is universal. This means that there is no such a problem domain where no causes and effects. The person can see nothing, but a cause or an effect exists.

A structure of cause-effect relations can form a causal chain. The causal chain begins with the first cause and follows with series of intermediate actions or events to some final effect. Though one link may not be as important or as strong like the other ones, they are all necessary to the chain. If just one of these intermediate causes is absent, then the final effect would not be reached. Additionally, even if you change something, you cannot remove the effect without removing or changing the cause.

2.2 Research Questions

Identification of cause-effect relations in a domain is a key element in the construction of the valid model of the domain. Sources of information about the domain can differ, and a human mind works using visual and audio information as well as its own background knowledge on the domain and the world. In case of automation of such analysis, most of information must be transformed into the verbal form to be able to use NLP tools.

NLP tools are able to perform the following tasks: tokenization, part-of-speech (POS) tagging, chunking, Name Entity Recognition (NER)/Classification, dependency analysis, constituency parsing, and coreference resolution. So, the result of the processing can be analysed further to identify causes and effects in the text.

The research questions are the following:

- RQ1: What natural language constructs for expressing cause-effect relations in text are used?
- RQ2: What models, patterns for identification of cause-effect relations in text are used?
- RQ3: What automatic techniques for extracting cause-effect relations from text are used?

The aim is to understand what natural language constructs may be ambiguous for NLP, what pitfalls exist in discovering cause-effect relations in text and what issues have been found in application of extracting tools.

Relevant part

3 CAUSE-EFFECT RELATIONS EXPRESSED IN TEXT

This section represents discussion on means for explicit and implicit expressing cause-effect relations in natural language, what patterns may indicate cause-effect relations in text at the sentence and discourse levels, as well as overview of research papers on automatic discovering cause-effect relations from text using NLP tools and other techniques.

3.1 Means for Expressing Cause-effect Relations in Text

Considering natural language processing and understanding tasks, researchers noted that causes and effects usually are states or events that can have different duration (Khoo et al., 2002; Solstad and Bott, 2017). The cause-effect relations in text may be expressed both explicitly and implicitly. And the very important aspect is similar language constructs used to express temporal and causal relations, and sometimes only temporal relations dictate the causal one (Ning et al., 2018).

3.1.1 Explicitly Expressed Relations

Several authors (Khoo et al., 2002; Solstad and Bott, 2017) mentioned that linguists have identified the following ways of explicitly expressing causes and effects:

- using *causal links* to link two phrases, clauses or sentences,
- using *causative verbs*,
- using *resultative* constructions,
- using *conditionals* (i.e., if…then constructions),
- using *causative adverbs*, *adjectives*, and *prepositions*.

As Khoo et al., (2002) stated, Altenberg (1984) had classified causal links into four main types:

- the adverbial link (e.g., *so*, *hence*, *therefore*). It can have a reference to the preceding clause or to the following clause;
- the prepositional link (e.g., *because of*, *on account of*). It connects a cause and an effect in the same clause;
- subordination. It can be expressed using *a subordinator* (e.g., *because*, *as*, *since*), *a structural link* marked by a non-finite *-ing* clause, and *a correlative comparative construction* (e.g., *so…that*);
- the clause-integrated link (e.g. *that's why*, *the result was*). Here they distinguish *thematic link*, when the linking words serve as a subject of the sentence, and *a rhematic link*, when the linking words serve as the complement of the verb.

One may say that causal links include as causal reasons as temporal reasons (Ning et al., 2018).

Causative verbs are "verbs the meaning of which include a causal element" (Khoo et al., 2002), e.g. "register" that in "X registers Y" means that "X causes Y to be *registered*". One of the working definitions of the causative verbs can be such that "Causative verbs specify the result of the action, whereas other action verbs specify the action but not the result of action" (Khoo et al., 2002).

A resultative construction (Khoo et al., 2002) is "a sentence in which the object of a verb is followed by a phrase decribing the state of the object as a result of the action denoted by the verb", e.g. "A user marked records yellow". A resultative phrase can be an adjective (the most common kind), a noun phrase, a prepositional phrase or a particle.

If-then conditionals ofthen indicate that the antecedent (the *if* part) causes the consequent (the *then* part). However, sometimes they just indicate a sequence of events not their cause-effect relation (Khoo et al., 2002).

Causative adverbs, adjectives and prepositions also can have a causal element in their meaning (Khoo et al., 2002). Causative adverbs can be different, the most interesting for us are adverbs that involve the notion of a result whose properties are context dependent (e.g. *successfully*), adverbs that refer not to causes but to effects (e.g. *consequentelly*), and adverbs of means (e.g. *mechanically*).

Causal adverbs and adjectives are not well studied (Khoo et al., 2002).

As Khoo et al., (2002) mentioned, prepositions also can be used to express causality. They can indicate a cause as proximity, a cause as a source, and a cause as volume.

3.1.2 Implicitly Expressed Relations

Implicit cause-effect relations usually are inferred by the reader using information in text and their *background knowledge* (Khoo *et al.* 2002; Solstad and Bott 2017; Ning *et al.* 2018). As Khoo et al. stated (Khoo *et al.* 2002) implicit causality can be inferred by several groups of verbs that "have "causal valence" – they tend to assign causal status to their subject or object". The authors referred to Corrigan's work (Corrigan 1993; Corrigan and Stevenson 1994), where the following groups of verbs had been identified:

- Experiential verbs (Experiencer-stimulus and Stimulus-Experiencer),
- Action verbs (Actor verbs and Non-actor verbs).

Experienced verbs describe someone having a particular psychological or mental experience. Therefore, experienced verbs can be skipped in the system analysis for software development.

Action verbs describe events. The subject of the verb can take the semantic role *agent or actor.* The object of the verb takes the role of *patient*. Some verbs give greater causal weight to the subject (*actor verbs*), other – to the object (*non-actor verbs*). At the moment, causal weigth seems not so important for domain analysis. However, the interesting thing is that both verbs have derived adjectives reffering to the subject or object. This means that some preceding actions can be expressed in text using not verbs, but adjectives. Some implicit causative verbs trigger expectations of explanations to occur in subsequent discourse (Solstad and Bott, 2017).

3.2 Identification of Cause-effect Relations in Text

Many theories exist for identification, modeling and analysis of cause-effect relations in psycholinguistics, linguistics, psychology and artificial intelligence. Those of theories attempting to reduce causal reasoning to a domain-general theory can be grouped as associative theories, logical theories and probabilistic theories (Waldmann and Hagmayer 2013).

Waldmann and Hangmayer (2013) stated that associative theories underestimate aspects of causality that are important for causal reasoning, however in some cases causes and effects can be identified only using associations.

Logical theories model causal reasoning as a special case of deductive reasoning. Waldmann and Hangmayer (2013) noted that *conditionals do not distinguish between causes and effects*, and background knowledge can be necessary to distinguish them as well as some temporal priorities.

Probabilistic theories considers causes as *"difference makers,* which raise (generative cause) or reduce (preventive cause) the probability of the effect*"* (Waldmann and Hagmayer 2013). However, as the authors noted, covariation does not necessarily reflect causation.

All the theories have their limitations in identification of causes and effects. In case of processing verbal (written) information to develop software, causes and effects mostly relate to business, mechanical and physical domains that certainly make a task easier for developers. At the present, logical theories seem to be the most suitable for this task and domains.

3.2.1 Verbal (Sentence) Domain

Solstad and Bott (2017) stated that verbs, as causative as action, indicate a cause between two events (3).

[[event1]] CAUSE [[event2]] (3)

v.

Where [[event1]] is when the subject does something, and [[event2]] is when the object changes its state. Besides that, it is inferred that the object wasn't in this state before the [[event1]] if otherwise is not mentioned in text.

Causing entities and the manner of the causing may be introduced using "by" phrases (Solstad and Bott, 2017).

As Khoo et al. mentioned (Khoo et al., 2002), the subject of the verb describing the event must be some *agent* or *actor*. It needs not be obligatory an animate *agent*, it may be an object, an abstract property, or an event (Solstad and Bott, 2017). Implicit causality verbs in most cases express causality between two *animate* objects followed by explanation (Solstad and Bott, 2017).

3.2.2 Discourse Domain

Solstad and Bott (2017) stated that causal relations such as explanations are crucial for understanding of discourse. Connections between causal relations may be expressed explicitly using causal links (Section 3.1) or implicitly. In the latter they must be inferred by the reader.

Some researchers (Kang et al., 2017; Solstad and Bott, 2017) indicate that at the level of discourse, the causal relations differ from thouse of at the sentence or clause level. At the discourse level they are supplemented with reasons and explanations. The authors assumed that the causal relations exist between entities that are propositional in nature: [[proposition1]] CAUSE [[proposition2]]. Sometimes, it is hard to understand are they express parallel or sequential propositions or explanations, as, for instance, in sentence "The user access was denied. The hackers taked the control."

Solstad and Bott stated that in case of implicit causality verb and discourse domains are mixed, where causal expressions connect causative verbs with reasoning and explanations within the same sentence (Solstad and Bott, 2017).

3.2.3 Conditionals

Conditionals do not express causal relations explicitly, but they involve causal models in their evaluation. Solstad and Bott (2017) mentioned that *If…then* constructs (i.e., *conditionals*) may form constructs hard for NLP analysis – the so-called *counterfactual* conditionals. They include such constructs as *might*, *would*, *if only*. They indicate possible "state of the world" in case of some "action" that would be done. For example, as in the sentence "If librarian would not have ordered the book, a manager assistant would have".

From one's viewpoint, such constructs must be avoided in the description of system functionality. However, counterfactual conditions may be used in expert systems to produce answers to queries of interest (Pearl, 2019).

3.2.4 Causal and Temporal Reasons

Ning et al., (2018) indicated that many of NLP research papers focus on the abovementioned language constructs (causative verb, causal links, discourse relations, etc.) skipping (or underestimating) temporal reasons. They believe that joint consideration of causal models and temporal models is more valuable for identifying and extracting cause-effect relations from text. The authors indicated that starting from 2016 researchers pay greater attention to this aspect (Mirza, 2014; Asghar, 2016; Mostafazadeh et al., 2016; Ning et al., 2018).

The interesting fact is that joint temporal and causal reasoning correctly identify counterfactual clauses (Ning et al., 2018).

3.3 Automated Extraction of Cause-effect Relations using NLP

Cause-effect relations are extracted using so-called linguistic and syntactic patterns that in most cases are created manually (at least at the beginning). Linguistic and syntactic patterns are based on means for explicit expressing causes and effects, e.g. causal links and causative verbs for linguistic patterns and verb phrases and noun phrases for syntactic patterns (Blanco et al., 2008; Ning et al., 2018; Mirza, 2014; Blass and Forbus, 2016).

Joint usage of both temporal reason model and causal model as well as Machine Learning (ML) are presented by several authors (Mirza, 2014; Ning et al., 2018).

The temporal model discovers a temporal relation between two events. The temporal relation can be annotated as *before*, *after*, *include*, *is_included*, *vague* (Ning et al., 2018), and *simultaneous*, *begins/ begun_by*, *ends/ended_by*, *during/during_inv*, *identity* (Mirza, 2014). Other authors (Mostafazadeh et al., 2016) use another annotations, i.e., *before*, *meets*, *overlaps*, *finishes*, *starts*, *contain* and *equals*. Their model distinguishes between *a precondition* and *a cause.* The causal models of all the mentioned authors discover causal relations between events using linguistic patterns. The authors state that analysis of both relation types allows extracting cause-effect relations even if they lack explicit causal reason.

A set of logical rules and Bayesian inference (in ML) are used by Sorgente, Vettigli and Mele (Sorgente et al., 2013; Sorgente et al., 2018). Bayesian inference is applied to exclude discovered cause-effect pairs that in essence are non-causal. Filtering takes into account such features as lexical features, semantics features (hyponyms and synonyms) and dependency features.

A comprehensive survey of automatic extraction of causal relations is presented by Asghar (2016). The author divided automatic methods into two groups:

- approaches that employ linguistic, syntactic and semantics patter matching, and
- techniques based on statistical methods and ML.

The first group started from small text fragments and evolved till large text corpuses. As Asghar stated the first group at their beginning was forced to prepare text fragments manually for automatic processing, like, for instance, in Blass' and Forbus' work (2016).

However, now this group uses NLP techniques to prepare cause-effect pairs (by using linguistic patterns) and then filtering them to reduce a number of non-causal pairs. Starting from the early 2000s, ML techniques have been used to gain extracting cause-effect relations. These techniques do not require a large set of manually predefined linguistic patterns, however, quality of learning depends on corpuses used for learning.

4 DISCUSSION

Summarizing results (Table 1), we can conclude that the larger number of overviewed research papers is focused on analysis of explicitly expressed cause-effect relations by using causal links and causative verbs (Sorgente et al., 2013; Sorgente et al., 2018; Asghar, 2016; Blanco et al., 2008; Mirza, 2014; Mostafazadeh et al., 2016). However, a few research papers pay their attention also to resultative constructions and causative prepositions. In other words, causal links, causative verbs and prepositions are more valuable for creation of linguistic/syntactic patterns for text processing. The advantage is small cost of preparation, while the result can be quite ambiguous.

Causative adverbs and adjectives up to 2018 are not well studied comparing to the main studies on causative verbs, causal links and temporal aspects of events and propositions.

According to survey in (Asghar, 2016), accuracy of results of extracting if…then conditionals is satisfactory only using ML techniques

Extracting multiple causes and effects is very domain-specific task, therefore only a few research solve it directly (Sorgente et al., 2013; Sorgente et al., 2018; Mueller and Hüttemann, 2018).

Cause-effect relations implicitly expressed by action verbs are analysed in the same group of causative verbs. While automated analysis of counterfactual conditionals is a quite hard task and some results are presented just in a few papers (Ning et al., 2018; Pearl, 2019).

Speaking about techniques used for automated cause-effect extracting, it could be found from results in Table 1 that preparation of text corpuses using predefined syntactic and linguistic patterns is less costly than using supervised ML techniques. However, the use of patterns limits discovered types of cause-effect relation only to these patterns. While ML enables discovering of much more cause-effect relations.

Filtering and statistical inferencing may be considered as a less expensive solution in comparison with ML techniques. However, some linguistic constructs may be ambiguous and, thus, results may differ from the desired ones. But statistical inferencing requires large datasets.

Ontology banks are also used, but moslty WordNet, VerbNet, PropBank and FrameNet are used sparsely.

The more successful results are shown by hybrid solutions where patterns, temporal reasons, ML and ontologies are presented. The limitation of the hybrid solutions is a lack of enough text corpuses for learning.

TFM construction requires processing verbal descriptions of the modelled environment. The descriptions contain information on system functioning within and interacting with its environment. Identification and extraction of causes and effects, as well as their relations, are vital for correct identification and specification of system's functional chacacteristics and causal dependencies between them.

Most of researh papers investigates cases with one cause and one (or two) effects. The TFM may have relationships between causal relations.

So, multi causes and multi effects must be idenified and extracted from text. However, there is just a few research papers presenting results on this, since this is a quite hard task.

It is clear that the starting point for extracting cause-effect relations from the descriptions of func-

Table 1: Automated extracting cause-effect relations using NLP and other techniques.

Lexical mark / discourse relation	Source	Pros and cons
	Explicit cause-effect relations	
Causal links: - the adverbial link - the prepositional link - subordination - the clause-integrated link	(Sorgente et al., 2013; Sorgente et al., 2018), survey in (Asghar, 2016); (Mirza, 2014; Blanco et al., 2008; Mostafazadeh et al., 2016)	Pros: small costs of preparation Cons: huge number of potential patterns, ambiguity
Causative verbs	(Sorgente et al., 2013; Sorgente et al., 2018), survey in (Asghar, 2016; Mostafazadeh et al., 2016)	Pros: small costs of preparation Cons: huge number of potential patterns, ambiguity
Resultative constructions	survey in (Asghar, 2016)	
Causative adverbs		
Causative adjectives		
Causative prepositions: - cause as proximity - cause as source - cause as volume	(Sorgente et al., 2013; Sorgente et al., 2018; Blanco et al., 2008)	Pros: small costs of preparation Cons: huge number of potential patterns, ambiguity
If-then conditionals	survey in (Asghar, 2016)	Cons: accuracy is satisfactory only using ML techniques
Multiple causes and effects (conjunctions)	(Sorgente et al., 2013; Sorgente et al., 2018; Mueller and Hüttemann, 2018)	
	Implicit cause-effect relations	
Counterfactual conditionals	(Ning et al., 2018; Pearl, 2019)	*Use for prediction*
Action verbs		*Consider as a subset of causative verbs*
	Techniques	
Temporal relations	(Mirza, 2014; Asghar, 2016; Mostafazadeh et al., 2016; Ning et al., 2018)	Pros: Analysis of event/proposition pairs where causality is implicit. Cons: Some linguistic constructs may be ambiguous.
Linguistic/syntactic patterns	(Ning et al., 2018), (Sorgente et al., 2013; Sorgente et al., 2018), survey (Asghar, 2016), (Blass and Forbus, 2016)	Pros: Does not require large corpuses of text for learning, domain-independent. Cons: Limited to the manually predefined set of patterns and propositions. A use of explicit causal indicators and in most cases ignoring implicit causalities.
Filtering (Bayesian inference, WordNet-based filtering, semantic filtering based on verb's senses)	(Sorgente et al., 2013; Sorgente et al., 2018), survey in (Asghar, 2016)	Pros: reducing a number of non-causal pairs. Cons: a set of pairs depends on a set of linguistic patterns
Machine Learning	survey in (Asghar, 2016), (Blanco et al., 2008; Mirza, 2014; Ning et al., 2018)	Pros: discovering implicit causality, analysis of ambiguous constructs, pre-conditions and postconditions Cons: requires large corpuses of text, may be domain-specific.
Statistical inferencing	survey in (Asghar, 2016)	Cons: requires large datasets
Ontology banks	survey in (Asghar, 2016), (Mostafazadeh et al., 2016; Kang et al., 2017)	Pros: WordNet is used frequently Cons: VerbNet, PropBank and FrameNet are used sparsely

tionality must be preparation of a corpus of linguistic/ syntactic patterns as well as more thorough analysis of if…then conditionals.

A use of temporal models, filtering and ontology banks seems more promising than a use of ML techniques since each problem domain will certainly have its own unique characteristics, but at the same time the diversity in description of functionality is not so defining than in descriptions of world phenomena.

5 CONCLUSIONS

The results of the survey show that identification of causes and effects as well as their relations can be based first on linguistic/syntactic patterns and temporal reason models. However, main disadvantage of the patterns is that it is not possible to identify all patterns for all types of cause-effect relations. The expression means of the natural language differ more than any set of predefined rules.

Besides that, one and the same pattern may be applied for both causal and non-causal relations. As well as not all linguistic and syntactic patterns have been researched, e.g. causal adverbs and adjectives. Filtering can help to solve this issue but is also limited to the known non-causal constructs. A use of temporal relation models is more valuable solution of this issue. However, some discourse descriptions may be very ambiguous, and there is no a guarantee that temporal relations will be identified correctly.

The more expensive and more flexible solutions are those of hybrid using machine learning, ontology banks and statistics. However, these solutions are more domain specific. They require a large amount of text corpuses for supervised learning of models. This could be a challenge, since not all the domains have them.

There are two clear trends in cause-effect relation extraction. The first is increasing the accuracy of the results using ontology banks, machine learning and statistical inferring. The second is decreasing the cost of these activities. The main challenge for construction of software models is a lack of a large amount of text corpuses and statistical datasets for potential problem domains. However, the potential positive aspect is that source documents for construction of software models may be limited to specifications (requirements, scenario, etc.) having less variability in expressing causality.

The future research direction is related to implementation of extracting causes and effects from the description of system functioning. The first step is to define a list of more frequent (potential) linguistic/syntactic patterns of causal dependencies in descriptions of system functioning. One of the very important aspects here is discovering of multi causes and multi effects. The accuracy of the obtained results will lead to the second step, i.e., to finding a solution that will show acceptable accuracy of results and will not be very expensive.

REFERENCES

Altenberg, B., 1984. Causal linking in spoken and written English. *Studia Linguistics*, 38(1), pp.20–69.

Asghar, N., 2016. Automatic Extraction of Causal Relations from Natural Language Texts: A Comprehensive Survey. *CoRR, abs/1605.0*. Available at: http://arxiv.org/abs/1605.07895.

Asnina, E., 2006. The Computation Independent Viewpoint: a Formal Method of Topological Functioning Model Constructing. *Applied computer systems*, 26, pp.21–32.

Asnina, E. and Ovchinnikova, V., 2015. Specification of Decision-making and Control Flow Branching in Topological Functioning Models of Systems. In *International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE), 2015*. Barcelona, Spain: SciTePress, pp. 364–373.

Blanco, E., Castell, N. and Moldovan, D., 2008. Causal Relation Extraction. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), pp. 28–30.

Blass, J.A. and Forbus, K.D., 2016. Natural Language Instruction for Analogical Reasoning: An Initial Report. In *Workshops Proceedings for the Twenty-fourth International Conference on Case-Based Reasoning (ICCBR 2016)*. pp. 21–30.

Bousetta, B., Beggar el, O. and Gadi, T., 2013. A methodology for CIM modelling and its transformation to PIM. *Journal of Information Engineering and Applications*, 3(2), pp.1–21.

Corrigan, R., 1993. Causal attributions to the states and events encoded by different types of verbs. *British Journal of Social Psychology*, 32, pp.335–348.

Corrigan, R. and Stevenson, C., 1994. Children's causal attribution to states and events described by different classes of verbs. *Cognitive Development*, 9, pp.235–256.

Elstermann, M. and Heuser, T., 2016. Automatic Tool Support Possibilities for the Text-Based S-BPM Process Modelling Methodology. In *Proceedings of the 8th International Conference on Subject-oriented Business Process Management - S-BPM '16*. New York, New York, USA: ACM Press, pp. 1–8.

Essebaa, I. and Chantit, S., 2016. Toward an automatic approach to get PIM level from CIM level using QVT rules. In *2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA)*. Mohammedia: IEEE, pp. 1–6.

Kang, D. et al., 2017. Detecting and Explaining Causes From Text For a Time Series Event. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. The Association for Computational Linguistics, pp. 2758–2768.

Kardoš, M. and Drozdová, M., 2010. Analytical method of CIM to PIM transformation in model driven architecture (MDA). *Journal of Information and Organizational Sciences*, 34(1), pp.89–99.

Khoo, C., Chan, S. and Niu, Y., 2002. The Many Facets of the Cause-Effect Relation. In R. Green, C. A. Bean, and S. H. Myaeng, eds. *The Semantics of Relationships: An Interdisciplinary Perspective*. Dordrecht: Springer Netherlands, pp. 51–70.

Kriouile, A., Addamssiri, N., Gadi, T. and Balouki, Y., 2014. Getting the static model of PIM from the CIM. In *2014 Third IEEE International Colloquium in Information Science and Technology (CIST)*. Tetouan: IEEE, pp. 168–173.

Kriouile, A., Addamssiri, N. and Gadi, T., 2015. An MDA Method for Automatic Transformation of Models from

CIM to PIM. *American Journal of Software Engineering and Applications*, 4(1), pp.1–14.

Kriouile, A., Gadi, T. and Balouki, Y., 2013. CIM to PIM Transformation: A Criteria Based Evaluation. *Int. J. Computer Technology and Applications*, 4(4), pp.616–625.

Miller, J. and Mukerji, J., 2001. Model Driven Architecture (MDA), Available at: http://www.omg.org/cgi-bin/doc?ormsc/2001-07-01.

Mirza, P., 2014. Extracting Temporal and Causal Relations between Events. In *Proceedings of the ACL 2014 Student Research Workshop*. Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 10–17.

Mostafazadeh, N., Grealish, A., Chambers, N., Allen, J. and 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, California: Association for Computational Linguistics, pp. 51–61.

Mueller, R. and Hüttemann, S., 2018. Extracting Causal Claims from Information Systems Papers with Natural Language Processing for Theory Ontology Learning. In *Proceedings of the 51st Hawaii International Conference on System Sciences (HICSS)*. Hawaii, USA: IEEE Computer Society Press.

Nazaruka, E., 2017. Meaning of Cause-and-effect Relations of the Topological Functioning Model in the UML Analysis Model. In *Proceedings of the 12th International Conference on Evaluation of Novel Approaches to Software Engineering*. SCITEPRESS - Science and Technology Publications, pp. 336–345.

Nazaruka, E., Ovchinnikova, V., Alksnis, G. and Sukovskis, U., 2016. Verification of BPMN Model Functional Completeness by using the Topological Functioning Model. In *Proceedings of the 11th International Conference on Evaluation of Novel Software Approaches to Software Engineering*. Portugal: SCITEPRESS - Science and Technology Publications, pp. 349–358.

Ning, Q., Feng, Z., Wu, H. and Roth, D., 2018. Joint Reasoning for Temporal and Causal Relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2278–2288.

Osis, J., 1969. Topological Model of System Functioning (in Russian). *Automatics and Computer Science, J. of Academia of Siences*, (6), pp.44–50.

Osis, J. and Asnina, E., 2011a. Is Modeling a Treatment for the Weakness of Software Engineering? In *Model-Driven Domain Analysis and Software Development*. Hershey, PA: IGI Global, pp. 1–14.

Osis, J. and Asnina, E., 2011b. Topological Modeling for Model-Driven Domain Analysis and Software Development: Functions and Architectures. In *Model-Driven Domain Analysis and Software Development: Architectures and Functions*. Hershey, PA: IGI Global, pp. 15–39.

Osis, J., Asnina, E. and Grave, A., 2007a. Formal computation independent model of the problem domain within the MDA. In J. Zendulka, ed. *Proceedings of the 10th International Conference on Information System Implementation and Modeling*, Hradec nad Moravicí, Czech Republic, April 23-25, 2007. Jan Stefan MARQ., pp. 47–54.

Osis, J., Asnina, E. and Grave, A., 2007b. MDA oriented computation independent modeling of the problem domain. In *Proceedings of the 2nd International Conference on Evaluation of Novel Approaches to Software Engineering - ENASE 2007*. Barcelona: INSTICC Press, pp. 66–71.

Osis, J., Asnina, E. and Grave, A., 2008. Formal Problem Domain Modeling within MDA. In J. Filipe et al., eds. *Software and Data Technologies: Second International Conference, ICSOFT/ENASE 2007, Barcelona, Spain, July 22-25, 2007, Revised Selected Papers*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 387–398.

Osis, J. and Donins, U., 2017. *Topological UML modeling: an improved approach for domain modeling and software development*, Elsevier.

Osis, J. and Slihte, A., 2010. Transforming Textual Use Cases to a Computation Independent Model. In J. Osis and O. Nikiforova, eds. *Model-Driven Architecture and Modeling-Driven Software Development: ENASE 2010, 2ndMDAandMTDD Whs*. SciTePress, pp. 33–42.

Pearl, J., 2019. The Seven Tools of Causal Inference, with Reflections on Machine Learning. *Communications of Association for Computing Machinery*, 62(3), pp.54–60.

Rhazali, Y., Hadi, Y. and Mouloudi, A., 2016. CIM to PIM Transformation in MDA: from Service-Oriented Business Models to Web-Based Design Models. *International Journal of Software Engineering and Its Applications*, 10(4), pp.125–142.

Šlihte, A. and Osis, J., 2014. The Integrated Domain Modeling: A Case Study. In *Databases and Information Systems: Proceedings of the 11th International Baltic Conference (DBandIS 2014)*. Tallinn: Tallinn University of Technology Press, pp. 465–470.

Slihte, A., Osis, J. and Donins, U., 2011. Knowledge Integration for Domain Modeling. In J. Osis and O. Nikiforova, eds. *Model-Driven Architecture and Modeling-Driven Software Development: ENASE 2011, 3rd Whs*. MDAandMDSD. SciTePress, pp. 46–56.

Solstad, T. and Bott, O., 2017. Causality and causal reasoning in natural language. In M. R. Waldmann, ed. *The Oxford Handbook of Causal Reasoning*. Oxford University Press.

Sorgente, A., Vettigli, G. and Mele, F., 2018. A Hybrid Approach for the Automatic Extraction of Causal Relations from Text. In *Emerging Ideas on Information Filtering and Retrieval*. Springer International Publishing AG, pp. 15–30.

Sorgente, A., Vettigli, G. and Mele, F., 2013. Automatic extraction of cause-effect relations in Natural Language Text. In C. Lai, G. Semeraro, and A. Giuliani, eds.

*Proceedings of the 7th International Workshop on Information Filtering and Retrieval co-located with the 13th Conference of the Italian Association for Artificial Intelligence (AI*IA 2013)*. pp. 37–48.

Waldmann, M.R. and Hagmayer, Y., 2013. Causal reasoning. In D. Reisberg, ed. *Oxford Handbook of Cognitive Psychology*. New York: Oxford University Press.