Towards Text Mining in Climate Science: Extraction of Quantitative Variables and their Relations

Erwin Marsi[1], Pinar Öztürk[1], Elias Aamot[1], Gleb Sizov[1], Murat V. Ardelan[2]

[1]Department of Computer and Information Science, [2]Department of Chemistry
Norwegian University of Science and Technology (NTNU)
{emarsi,pinar,sizov}@idi.ntnu.no, eliasaa@stud.ntnu.no, murat.v.ardelan@ntnu.no

Abstract

This paper addresses text mining in the cross-disciplinary fields of climate science, marine science and environmental science. It is motivated by the desire for literature-based knowledge discovery from scientific publications. The particular goal is to automatically extract relations between quantitative variables from raw text. This results in rules of the form "If variable X increases, than variable Y decreases". As a first step in this direction, an annotation scheme is proposed to capture the events of interest – those of change, cause, correlation and feedback – and the entities involved in them, quantitative variables. Its purpose is to serve as an intermediary step in the process of rule extraction. It is shown that the desired rules can indeed be automatically extracted from annotated text. A number of open challenges are discussed, including automatic annotation, normalisation of variables, reasoning with rules in combination with domain knowledge and the need for meta-knowledge regarding context of use.

Keywords: Text Mining, Literature-based Discovery, Climate Science, Marine Science, Environmental Science, Corpus Annotation, Relation Extraction, Event Extraction

1. Introduction

One of the characteristics of the cross-disciplinary fields of climate science, marine science and environmental science is the existence of many different processes that affect each other in direct and indirect ways, resulting in highly complex systems. In climate science, for example, *climate feedback* is defined as "An interaction in which a perturbation in one climate quantity causes a change in a second, and the change in the second quantity ultimately leads to an additional change in the first. A negative feedback is one in which the initial perturbation is weakened by the changes it causes; a positive feedback is one in which the initial perturbation is enhanced." (Stocker et al., 2013). Identifying such feedback processes is generally considered a crucial step in understanding and predicting phenomena like global warming.

Unfortunately much potential knowledge regarding processes and their interactions is hidden in the scientific literature, scattered over journals catering to different scientific communities with relatively little communication among them. Given the vast and constantly growing nature of this body of literature, it is indeed hard for individual researchers to keep track of all relevant publications in their field of expertise, let alone of those in related or even more distant areas.

Text mining of scientific literature may contribute to alleviating this problem (Etzioni, 2011). Climate, marine and environmental science may all benefit from automatic extraction of processes and their interactions from scientific publications. Extracted information can be indexed, thus allowing researchers to search for interactions between processes in a much more effective way than with conventional keyword-based search engines.

In addition, this structured data can be used for inference in discovery support systems. For example, pairs of cause and effect processes can be chained together, possibly in combination with existing domain knowledge, in order to suggest hypotheses about indirect interactions or feedback loops or to point out contradictory findings. Such discovery of implicit knowledge in a body of literature is aimed for in the field of *literature-based discovery* (LBD). The first results in LBD were produced by Swanson (1986a) through manually executing a search algorithm based on co-occurrence statistics of terms. This allowed him to combine two publicly available knowledge fragments – (1) Fish oils reduce blood viscosity and (2) patients with Raynaud's disease tend to exhibit high blood viscosity – to form the hypothesis that fish oils treat Raynaud's disease. The hypothesis was later confirmed experimentally. Even though both knowledge fragments were publicly available to any researcher, nobody had been aware of both knowledge fragments and made the connection.

The aim of LBD is to create systems that provide discovery support to uncover such potential hypotheses, which Swanson referred to as *undiscovered public knowledge* (Swanson, 1986b). Most LBD methods are based on chaining of unspecified relations, using term co-occurrence frequencies as heuristic evidence for a relation between two terms. Terms are usually extracted as n-grams from the text (Lindsay and Gordon, 1999) or taken from a controlled vocabulary or ontology (Weeber et al., 2001).

Recently, co-occurrence based LBD methods have come under critique for yielding imprecise results, as they fail to exploit the true breadth of knowledge contained in the scientific literature. Hristovski et al. (2008) therefore advocate a text mining based approach, where relation extraction is used to discover specific relations between two concepts. This enables more precise and complex query patterns.

LBD efforts along these lines in the biomedical domain have taken advantage of existing tools such as SemRep

(Rindflesh and Fiszman, 2003). SemRep is a major text mining system for the biomedical domain that exploits structured domain knowledge found in the Unified Medical Language System (UMLS)[1]. The UMLS consists of three tools: a lexicon, a semantic network and a meta-thesaurus. Using underspecified symbolic language processing, SemRep is able to extract a wide range of specific relations, such as TREATS, HAS_PART and LOCATION_OF. However, adapting such tools to less resourced domains, such as marine science, is difficult because of the lack of resources like UMLS and because of the knowledge, time and effort required for writing extraction rules.

Other work has concentrated on relation and event extraction through machine learning. Causal relations are one of the most commonly targeted relations. In (Mihăilă and Ananiadou, 2013), several machine learning algorithms are applied to recognise causality triggers such as *therefore*, *because*, *as a result of*, etc. The approach is tested on BioCause (Mihaila et al., 2013) and BioDRB (Prasad et al., 2011) corpora, which consist of articles with manually annotated causal relations between named entities in the biomedical domain. A machine learning approach has also been applied by Pechsiri and Piriyakul (2010) to extract causal relations in the agricultural domain, which are then used to construct explanation knowledge graphs that represent the domain knowledge. Supervised learning requires domain-specific training material though, which is currently lacking in our domain of climate science, marine science and environmental science.

An alternative approach may be the use of unsupervised learning and clustering techniques. As an example of unsupervised techniques for causal relation extraction, Hashimoto et al. (2012) propose a set of relations that can be used to detect causality. They identify excitatory, inhibitory and neutral relations with a corresponding set of extraction templates. More templates are acquired automatically by a bootstrapping process. Excitatory relations are then used for extraction of contradictions, causality relations and generation of causality hypotheses. A disadvantage of these approaches is that their performance is generally far less accurate than that of supervised methods. In addition, it seems they are not capable of covering the more complex events of changing processes and their interactions, as we are interested in here.

Other approaches have explored extraction of more fine-grained types of events, including those of increase and decrease. Zambach and Lassen (2010) identify and linguistically analyse verbs that express regulation relations, positive and negative, between processes and substances in the biomedical domain. They suggest that their analysis can benefit extraction of, as well as reasoning over, these relations in the biomedical domain, although no implementation or evaluation was carried out.

In sum, currently text mining in climate science, marine science and environmental science appears to be virtually non-existent. Our research agenda targets developing text mining in this area, in particular towards applications in LBD. Existing approaches and tools from other domains such as biomedicine are not readily applicable and do not provide extraction of the type of processes and interactions needed for our purposes. Development of domain-specific event extraction tools is therefore high our agenda. Following the lead of text mining initiatives in biomedicine (Kim et al., 2009), we explore manual text annotation for creating annotated corpora, which can be used to train classifiers for automatic annotation, and ultimately automatic rule extraction. In Section 2., an annotation scheme is proposed to annotate the events of interest – those of change, cause, correlation and feedback – as well as the entities involved in them. Its purpose is to serve as an intermediary step in the process of rule extraction. It is shown in Section 3. that such rules can indeed be automatically extracted from annotated text. Section 4. discusses a range of open challenges, including automatic annotation, normalisation of variables, reasoning with rules in combination with domain knowledge and the need for meta-knowledge regarding context of use. The final Section 5. lists conclusions and future work.

2. Annotation

2.1. Procedure

The data consisted of 12 abstracts (2369 words) from recent, high-quality scientific journal publications about the relation between climate and ocean changes. These were selected by our domain expert (author MVA) as a reasonably representative sample of the text type in the targeted area, comprising multi-disciplinary work in marine biology, marine chemistry, oceanography, environmental science, climate science, biogeoscience and geophysics. Text was automatically extracted from PDF files. Abstracts were manually extracted, tokenised and split into sentences, also allowing for manual correction of minor PDF-to-text conversion errors.

Annotation was carried out using the Brat annotation tool (Stenetorp et al., 2012). The annotation scheme described below was developed in an iterative fashion in close collaboration with our domain expert. It is inspired by annotation efforts in the biomedical domain such as the GENIA corpus (Kim et al., 2003) and the corpora used in the BioNLP shared tasks on event extraction (Kim et al., 2009). It covers a particular type of events – those of change, cause, correlation and feedback – and the entities involved in them, quantitative variables. The primary reason for annotation is not to analyse the text according to some linguistic formalism or theory, or to follow some knowledge representation formalism or ontological theory. Instead the purpose of the annotation is rather pragmatic: to serve as an intermediary step in the process of extracting rules about the relation between quantitative variables from raw text.

2.2. Annotation scheme

The resulting annotation scheme involves one type of entity (variable), several types of events (change, increase, decrease, cause, correlate, feedback) and some basic logic structure (and/or, negation).

2.2.1. Variables

A quantitative variable is an entity that can be counted or measured. Its value can be naturally expressed by a number

[1] http://www.nlm.nih.gov/research/umls/

such as a count, a ratio, a percentage or a scalar (quantity of units). It can be regarded as a (potential) quantitative variable in an experiment or a model. Not every variable in the text is labeled as such. To save annotation time and effort, only those variables related to a change are annotated. The direction of change can be positive (increasing), negative (decreasing) or unspecified (either increasing or decreasing), but there must always be a clear cue in the text that the variable is involved in some change. Examples of changing, increasing and decreasing variables respectively:

(1) a. significant changes in [*surface ocean pH*]
b. rise in [*atmospheric CO2 levels*]
c. decline in [*marine primary production*][2]

In contrast, the text spans in (2) are not annotated as variables.

(2) a. *[*carbon dioxide*] and [*light*] are two major prerequisites of photosynthesis
b. *changes in [*the network of global biogeochemical cycles*]
c. *The concentrations of [*DFe*] and [*TaLFe*] were relatively high

The text spans in (2-a) are measurable, in principle at least, but there is no textual cue in the context indicating that they are subject to change. The text span in (2-b) admittedly identifies something that is changing, but it is an abstraction – not something that can be measured and naturally expressed through a number. The ones in (2-c) express a static state rather than a dynamic event. The reason for excluding cases like these is that they do not lead to useful rules about the relation between quantitative variables.
Variables must be indicated as precisely as possible, that is,including any relevant specifications, modifications or conditions. So instead of (3-a), (3-b) is preferred.

(3) a. *a difference in [*carbon concentration*] between the ocean surface and the deep waters
b. a difference in [*carbon concentration between the ocean surface and the deep waters*]

The choice is motivated by the assumption that, given a syntactic parse, it is usually easier to generalize a complex argument by stripping modifiers than the other way around. Variables are tagged with the label VARIABLE. We intend to distinguish different subclasses of variables, resulting in a more fine-graned categorisation of entities, in the near future. For now, we focus on annotation of the events and basic logic structure.

2.2.2. Change, Increase and Decrease

A change is an event in which the value of a quantitative variable is changing. The direction of change can be positive (increasing), negative (decreasing) or unspecified (either increasing or decreasing), but there must always be a clear cue in the text that the variable is involved in a change. This is referred to as the *trigger* for the event.

Examples of triggers for event types of change, increase and decrease are:

(4) a. [*regional changes in*] phytoplankton
b. [*addition of*] labile dissolved organic carbon
c. [*to slow down*] calcification in corals

Changes must apply to a variable; hence the text span in (5) does not trigger a change event.

(5) *marine primary production is sensitive to climate [*variability and change*]

Events of increase, decrease and undirected changes are tagged as INCREASE, DECREASE and CHANGE respectively. Events are related to variables through thematic roles, which specify the different participants in the event. Change events must always have a THEME role that is filled by the variable that is changing. Typical annotation examples are therefore:[3]

(6) a. [DECREASE reduced] [THEME calcite production]
b. [CHANGE significant changes in] [THEME surface ocean pH]

Change events can also function as Cause/Correlate events, as will be described in the next Section, in which case they take an AGENT or CO-THEME role as well.

2.2.3. Cause

Cause events involve a pair of changes where the first change causes the second change. Since a change event involves a changing variable, as its theme, causal events thus express a causal relation between two changing variables. The trigger of a cause event is annotated with a CAUSE tag. Triggers are often verbs, but can also be adjectives (*stimulatory*), adverbs (*therefore*) or subjunctive phrases (*due to*, *in response to*) or other phrasal expressions (*has an effect on*).
Cause events must always have two thematic roles: an AGENT identifying the cause and a THEME identifying the effect. Examples of cause events are:

(7) a. [AGENT rise in atmospheric CO2 levels] [CAUSE *causes*] [THEME significant changes in surface ocean pH]
b. [AGENT Fe(III) addition in the presence of GA (FeGA)] [CAUSE *gave*] [THEME higher Fe(II) concentration]
c. [AGENT diminished calcification] [CAUSE *led to*] [THEME a reduction in the ratio of calcite precipitation to organic matter production]

In many cases, a cause event and a change event share one and the same trigger, as in the following examples:

(8) a. [AGENT changes in the magnitude of total and export production] [CHANGE *can strongly influence*] [THEME atmospheric CO2 levels]
b. [THEME calcification and net primary production] [INCREASE *are significantly increased by*]

[2]The total amount of energy produced by marine organisms such as photosynthetic plankton.

[3]We use labeled brackets to denote entities, events or thematic roles, depending on the context of discussion.

[AGENT high CO2 partial pressures]

c. [AGENT addition of labile dissolved organic carbon] [DECREASE *reduced*] [THEME phytoplankton biomass]

In (8-a), *can strongly influence* serves as the cue for a change event with variable *atmospheric CO2 levels* as its theme. At the same time, it is the trigger for a cause event with agent *changes in the magnitude of total and export production* and theme *atmospheric CO2 levels*. In principle, both events can be annotated separately. However, in order to avoid a needlessly complex annotation, we chose to not annotate the cause event explicitly. Instead *changes in the magnitude of total and export production* is given the role of agent in the change event. Presence of an agent role suffices to infer the cause event. In other words, where there is both a cause event and a change event, we annotate the change event because it can be inferred from the presence of the agent role that a cause event is also being annotated. Two more instances of this pattern are shown in (8-b)[4] and (8-c).

2.2.4. Correlate

Correlate events involve a pair of changes where the first change correlates with the second change. Since a change event involves a changing variable, as its theme, correlate events thus express a correlation between two changing variables. That is, if one of them changes, the other changes along. Correlations have two roles, THEME and CO-THEME, both of which should be fulfilled by a change event (i.e. INCREASE, DECREASE or CHANGE). Examples of correlate events:

(9) a. [THEME reduced calcite production] [CORRELATE *was accompanied by*] [CO-THEME an increased proportion of malformed coccoliths]

b. [THEME carbon:nutrient ratio turns out to decrease] [CORRELATE *with*] [CO-THEME increasing mixed-layer depth and temperature]

c. Here we report [THEMe reduced calcite production] [CORRELATE *at*] [CO-THEME increased CO2 concentrations]

d. [CORRELATE *When*] [CO-THEME bacterial growth rate was limited by mineral nutrients], [THEME extra organic carbon accumulated in the system]

Notice that correlation can be triggered by a verb (9-a), a preposition (9-b-c) or an adverb/conjunction (9-d).
Statistically speaking, correlation is not a directional relation, in contrast to causation. That is, if a change in variable A is correlated with a change in variable B, then it follows that a change in variable B is correlated with a change in variable A. However, in discourse there is often a distinction between a variable of interest (the dependent variable) and a related variable (the independent variable). Thus even though strictly speaking there is no causal relation between the two variables, the text usually takes a particular perspective, suggesting one is more central than the other. By convention, the central variable is tagged as THEME, whereas the other one is tagged as CO-THEME. The rule of thumb is that the co-theme is syntactically the argument of a preposition (e.g. *with, at, under*) or an adverb/conjunction (e.g. *when*).

Occasionally correlations can hold between a change event and a variable, or even between two variables, rather than between two change events. In these exceptional cases, we assume the variable is interpreted as changing (i.e. as being part of an implicit change event), because it is involved in a correlate event. Two examples of this exceptional pattern are:

(10) a. [THEME:INCREASE Concentrations of DFe increased slightly] [CORRELATE *with*] [CO-THEME:VARIABLE depth in the water column]

b. [THEME:VARIABLE growth rates in the high-CO2-grown cells] [CORRELATE *were related to*] [CO-THEME:VARIABLE light level]

In (10-a), the role of co-theme is not taken by a change event, but by the variable *depth in the water column*. It is thus assumed that the depth in the water column is a changing variable in the correlation described. Similarly, (10-b) has both roles of the correlate event taken up by variables, which are therefore interpreted as subject to change.

2.2.5. Feedback

Feedback loops are an important concept in climate science. An example is that of the relation between rising temperature and methane release: a rise in temperature causes more permafrost to melt, which causes more release of methane in the atmosphere (a "green house" gas), which causes further rising of the temperature, and so on. However, explicit mentioning of feedback events in the text appears to be rare compared with the frequent occurrence of change events, so our proposal for annotation of feedbacks is currently based on only a couple of instances. Feedback events hold between two variables, filling the roles of THEME and CO-THEME, as exemplified below:

(11) our model suggests the existence of [+FEEDBACK *a positive feedback between*] [THEME temperature] and [CO-THEME atmospheric CO2 content]

Analogously to change events, feedback events can be positive (self-sustaining, self-enhancing), negative (self-stabilising, self-diminishing) or of unspecified polarity. Positive or negative feedback are annotated with an attribute whenever a trigger is present.

2.2.6. Referring expressions

Referring expressions such as anaphoric expressions (e.g. *it*, *this*) and underspecified definite descriptions (e.g. *the process*) are annotated only in so far as they play a thematic role in an event of interest. Consider the following narrative:

[4] The agent in example (8-b), i.e. *high CO2 partial pressures*, is arguably not an event but a state. However, we took the liberty to interpret this as *increasing CO2 partial pressures* in this context, which is in accordance with the interpretation of our domain expert.

(12) s1: Future shoaling of upper-mixed-layer depths will expose phytoplankton to [INCREASE increased] [THEME mean light intensities].
s2: [REFEXP/AGENT *This*] [CAUSE may cause] [DECREASE a widespread decline in] [THEME marine primary production]

A graphical representation of a slightly extended version of this example is shown in Figure 1. The first sentence contains an INCREASE event, which is referred to in the second sentence by means of the referring expression *This*, establishing it as the cause for the DECREASE event. Such referring expressions must therefore be resolved in order to deduce the rule that an increase in mean light intensities causes a decrease in marine primary production. In order to achieve this, they are tagged as REFEXP and connected with their antecedent by means of a COREF relation.

2.2.7. Combinations

Variables or events can be combined through conjunction or disjunction. Such combinations are labeled as AND or OR, where their constituents fill the role of PART. In (13-a), for example, the combination AND serves as the theme of the INCREASE event. Likewise, two increasing events are combined to serve as the theme in a causal event.

(13) a. [INCREASE increasing] [PART:VARIABLE mixed-layer depth] [THEME:AND and] [PART:VARIABLE temperature]
b. [CAUSE gave] [PART:INCREASE higher] [THEME Fe(II) concentration] [THEME:AND and] [PART:INCREASE higher] [THEME growth rate of phytoplankton]

The alternative option in (13-a) is to tag the whole combined phrase as a single variable. We chose not do so because coordination is a notoriously hard problem for syntactic parsers and any help from the annotation in resolving ambiguity should be exploited. Notice also that a similar option is not available in (13-b), as considering the whole combination as a single change event would result in loss of substantial information.

There are certain cases, like where an adjectival modifier modifies a conjunction of two variables, that can not be accommodated by the proposed annotation scheme. This is not a shortcoming of the Brat annotation tool, but a matter of trade-off between expressivity and complexity: covering these instances requires additional relations or events, which would further complicate the annotation process. However, judging from the sample texts annotated so far, these cases are rare.

2.2.8. Negation

Events can carry a negation attribute to account for examples such as:

(14) a. TaLFe [CORRELATE+NEG did *not* show any consistent trend with] depth
b. [CHANGE+NEG *No* differences] in cellular organic carbon:nitrogen ratios were observed

Triggers for negation are currently not explicitly annotated.

3. Rule extraction

The proposed annotation allows for automatic extraction of rules about the relations between quantitative variables. There are three main types of rules: causal rules, correlation rules and feedback rules.

Causal rules are of the type "If variable X changes, than variable Y changes". An example of such a rule and its source text is shown in Figure 1. The notation uses single arrows to denote changing variables, where '$\uparrow$' stands for 'increasing', '$\downarrow$' for 'decreasing' and '$\updownarrow$' for 'changing.' Parts of a combination are joined by '$\wedge$' or '$\vee$' and delimited by square brackets. A causal relation is denoted by the double arrow '$\Longrightarrow$'. Causal rules are basically extracted by looking for CAUSE events, taking their AGENT and THEME roles for cause and effect respectively. Notice that in Figure 1, interpreting combinations and resolving referring expressions to their antecedent takes some additional processing. Another source for causal rules is change events with both AGENT and THEME roles.

Correlation rules are of the type "Changes in variable X correlate with changes in variable Y", as exemplified in Figure 2. The curly arrow '$\rightsquigarrow$' is used to indicate the relation between a independent and a dependent variable. These rules are extracted from CORRELATE events, using their CO-THEME role as the independent variable (LHS of the rule) and their THEME role as the dependent variable (RHS of the rule).

Feedback rules, an example of which is shown in Figure 3, are of the form: "Changes in variable X feed back through changes in variable Y". The feedback relation is denoted by a double sided arrow '$\Longleftrightarrow$', optionally with a superscripted '+' or '-' for positive and negative feedback respectively.

Notice that conceptually a feedback relation is assumed to hold between changing events. However, often there is no explicit trigger for a change event present in the text. For example, in the annotation in Figure 3, both roles are filled by variables instead of change events. Such variables are therefore 'promoted' to change events during rule extraction, resulting in '$\updownarrow$ temperature' and '$\updownarrow$ marine primary production'. Similar promotions apply occasionally to variables in events of change, cause or correlation (cf. Section 2.2.4.).

4. Discussion

The annotation scheme proposed below seems a good candidate for the purpose of rule extraction. However, it has only been tried on a small set of abstracts and it remains to be seen how it holds up when applied to more text. To provide some indicative statistics, the pilot-corpus contains the following number of labels: 107 VARIABLE, 33 CHANGE, 82 INCREASE, 50 DECREASE, 20 CAUSE, 26 CORRELATE, 32 AND, 2 OR, 5 REFEXP and 2 NEGATION. Annotation of more text is required to settle certain corner cases and is likely to reveal additional issues. For example, the current scheme can not capture the fact that *ocean acidification* is an event, i.e., a decrease of the pH of the ocean water. If similar examples turn out to occur frequently, this may cause a revision of the annotation scheme. Inter-annotator agreement has not been measured so far. In addition to this,

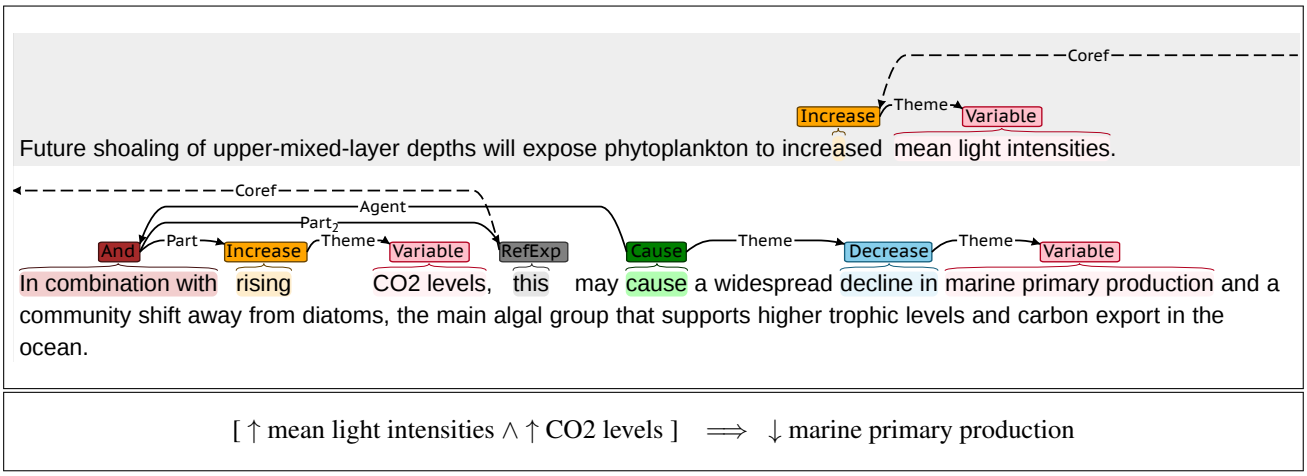

Figure 1: Example of a causal rule extracted from a pair of annotated sentences

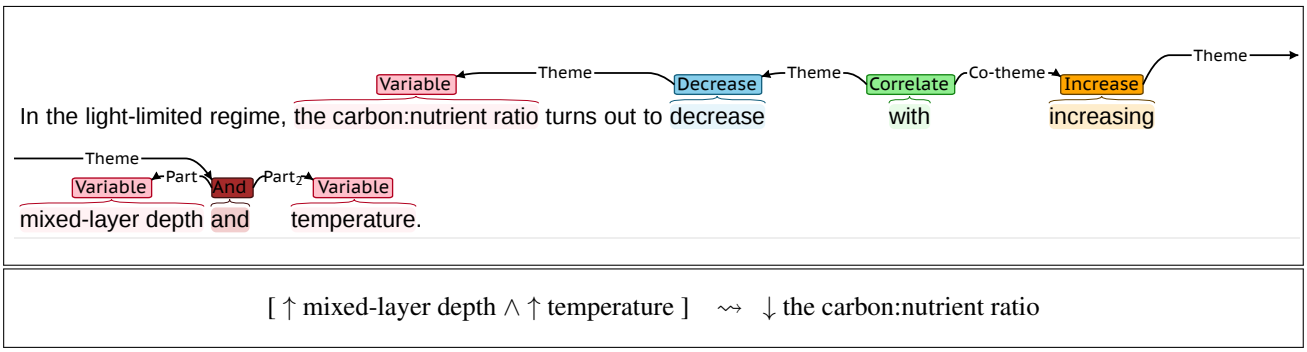

Figure 2: Example of a correlation rule extracted from an annotated sentence

we are also considering a type of evaluation in which extracted rules and their corresponding source texts are shown to domain experts, who are then asked to judge if the rule is entailed by the text.

Manual annotation is costly. There are at least two strategies which may reduce annotation time and costs. The first one is to bootstrap from existing extraction systems. Recent advances in *open information extraction*, where there is no predefined set of entities and relations, have resulted in open source systems like ReVerb (Fader et al., 2011) and it successor OpenIE. Banko and Etzioni (2008) claim that when the number of target relations is small, and their names are known in advance, an open IE system is able to match the precision of a traditional supervised extraction system, though at substantially lower recall. This suggests that at least a part of the annotation can be accelerated with the help of such tools.

A second strategy to reduce annotation costs involves the use of *active learning*, which is a training method for supervised learners that tries to obtain maximal performance gain with minimal annotation effort (Olsson, 2009). It is an iterative procedure, starting with a small amount of labeled data and a large amount of unlabelled data. In each iteration, a classifier is trained on the labeled data and subsequently applied to the unlabelled data. Only the most informative instances – e.g., those for which classification confidence is lowest – are passed on to a human annotator for manual annotation. These manually labeled instances are added to the training data and the procedure is repeated. Good results have been reported with the use of active learning, e.g. by (Gambäck et al., 2011).

The extracted rules expressing relations of correlation, causality or feedback between quantitative variables are intended to be used in knowledge discovery support systems. One use case is to search for other variables directly related to a certain variable of interest. For example, find all processes that affect or are affected by a rise in atmospheric CO2 level. The variable in question may be expressed in many different ways though, for example, as *CO2*, *atmospheric CO2*, *CO2 concentrations* or *CO2 partial pressures*, but not as *CO2 levels in oceanic surface waters* or *the distribution of CO2 between the atmosphere and the ocean*. Simple string matching between the variables in queries to those in rules will given limited recall and precision. Related to this is the issue of differences in terminology across research fields. For instance, *export production* and *biological pump* are different terms, used by chemists and biologists respectively, for the same process of carbon cycling in the oceans. One possible strategy to cope with this issue is to have a more fine-grained categorisation of entities, allowing different surface realisations to be mapped to the same underlying domain concept. This would allow more general rules to be extracted, and could also be beneficial in helping to bootstrap lexical resources.

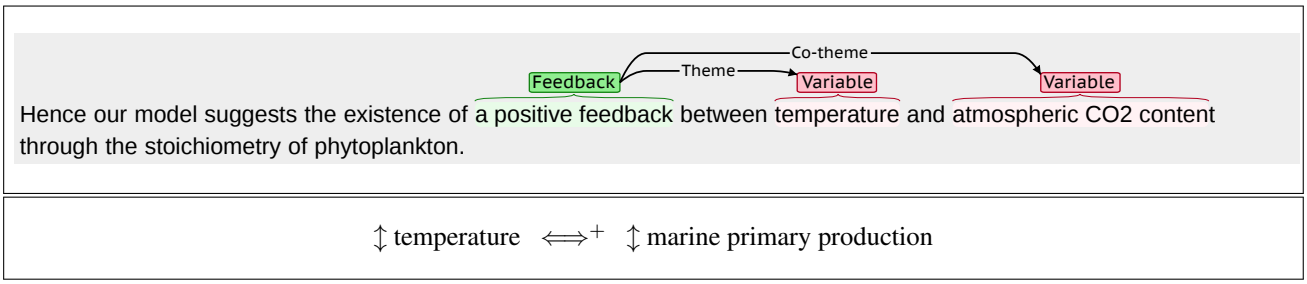

Figure 3: Example of a feedback rule extracted from an annotated sentence

Ultimately all relevant entities may be normalised by linking them to a unique concept in a domain ontology (Bada et al., 2012). However, whereas the concepts of interest in biomedicine are relatively well understood – including such entities as cells, proteins and genes – and covered by widely used ontologies, such common ground currently seems to lack in climate, marine and environmental science science.

A different but related problem is exemplified in correlation rule (15-b) extracted from the second part of sentence (15-a).

(15) a. Concentrations of DFe increased slightly with depth in the water column, while that of TaLFe did not show any consistent trend with depth.

b. ¬ [↕ depth ⇝ ↕ that of TaLFe]

The problem is that *depth* in (15-b) is too general and should in fact be linked to *depth in the water column* for proper interpretation. Likewise, *that of TaLFe* should be interpreted as *concentrations of TalFE*. This illustrates the need for coreference resolution and more general, linking of subsequent mentions of the same entity in the text, a notoriously hard task in NLP.

Apart from search, another use case for extracted rules is to generate potential hypotheses about indirect relations between variables or feedback loops among them. This can be accomplished by chaining together two or more rules, matching the change event on the right-hand-side of one rule to a similar change event on the left-hand-side of another rule. Matching gives rise to the same problems discussed above, i.e., different ways of referring to the same entity. In addition, there is the issue of context-dependency. Most rules are not universally applicable, but only apply under certain conditions in a particular context. For example, a rule may be limited in scope to certain biological species or organisms, a particular geographical region or historical time period, subject to a given assumption (*only if* ...), etc. This is related to initiatives for annotating meta-knowledge such as confidence level (fact vs. conjecture), source (resulting from observation vs. analysis) or origin (present or cited work) as in (Thompson et al., 2011). Proper modelling of rule context would require a rather deep understanding of the whole text. Although we acknowledge the importance of conditions on events, we intend to leave their annotation to a later stage. For now, we plan to leave this to the user by offering facilities in the user interface to quickly inspect the source text for each rule.

Inference with rules may be further enhanced by exploiting domain knowledge. For example, given an ontology which contains the fact that *diatoms* are a kind of *phytoplankton*, rules containing either of the terms may be generalised by substituting the hypernym or specialised by substituting the hyponym. In a similar vein, rules can be generalised by removing specifiers, modifiers or parts of a conjunction. Whether or not this constitutes valid inference seems connected to recent developments in textual entailment, in particular work on natural logic (MacCartney and Manning, 2008).

5. Conclusion

An annotation scheme was proposed to capture events of change, cause, correlation and feedback, as well as the entities involved in them, in the cross-disciplinary fields of climate science, marine science and environmental science. It was shown that rules about the relation between changing processes can be automatically extracted from annotated text. Follow-up work will involve annotating more text, as well as measuring inter-annotator agreement and rule adequacy. Simultaneously, tools for automatic annotation will be developed. Future work will also address normalisation of entities, tracking of entity mentions, modelling of rule context and combination with domain knowledge.

6. Acknowledgements

Financial aid from the European Commission (OCEAN-CERTAIN, FP7-ENV-2013-6.1-1; no: 603773) is gratefully acknowledged. We thank the reviewers for their valuable comments.

7. References

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner, K. Bretonnel Cohen, Karin Verspoor, Judith A. Blake, and Lawrence E. Hunter. 2012. Concept annotation in the CRAFT corpus. *BMC bioinformatics*, 13(1):161+, July.

Michele Banko and Oren Etzioni. 2008. The tradeoffs between open and traditional relation extraction. In *Proceedings of ACL-08: HLT*, pages 28–36, Columbus, Ohio, June. Association for Computational Linguistics.

Oren Etzioni. 2011. Search needs a shake-up. *Nature*, 476(7358):25–26, August.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Björn Gambäck, Fredrik Olsson, and Oscar Täckström. 2011. Active learning for dialogue act classification. In *Twelfth Annual Conference of the International Speech Communication Association*.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong H. Oh, and Jun'ichi Kazama. 2012. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 619–630, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dimitar Hristovski, C. Friedman, T. C. Rindflesch, and B. Peterlin. 2008. Literature-Based Knowledge Discovery using Natural Language Processing. In Peter Bruza and Marc Weeber, editors, *Literature-based Discovery*, volume 15 of *Information Science and Knowledge Management*, chapter 9, pages 133–152. Springer, Heidelberg, Germany.

J. D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpusa semantically annotated corpus for biotextmining. *Bioinformatics*, 19(suppl 1):i180–i182, July.

Jin D. Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, Colorado, June. Association for Computational Linguistics.

Robert K. Lindsay and Michael D. Gordon. 1999. Literature-based discovery by lexical statistics. *Journal of the American Society for Information Science*, pages 574–587.

Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 521–528, Stroudsburg, PA, USA. Association for Computational Linguistics.

Claudiu Mihăilă and Sophia Ananiadou. 2013. Recognising discourse causality triggers in the biomedical domain. *Journal of bioinformatics and computational biology*, 11(06).

Claudiu Mihaila, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14(1):2+.

F. Olsson. 2009. A literature survey of active machine learning in the context of natural language processing. Technical Report Tech. Rep. T2009, SICS, Stockholm, Sweden.

Chaveevan Pechsiri and Rapepun Piriyakul. 2010. Explanation knowledge graph construction through causality extraction from texts. *Journal of Computer Science and Technology*, 25(5):1055–1070.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. 2011. The biomedical discourse relation bank. *BMC bioinformatics*, 12(1):188+, May.

Thomas C. Rindflesh and Marcelo Fiszman. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. *J. of Biomedical Informatics*, 36(6):462–477, December.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA, April. Association for Computational Linguistics.

Thomas F Stocker, Q Dahe, and Gian-Kasper Plattner. 2013. Climate change 2013: The physical science basis. *Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Summary for Policymakers (IPCC, 2013)*.

Don R. Swanson. 1986a. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18.

Don R. Swanson. 1986b. Undiscovered public knowledge. *The Library Quarterly*, 56(2):pp. 103–118.

Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. 2011. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12(1):393+, October.

Marc Weeber, Henny Klein, Lolkje T. de Jong van den Berg, and Rein Vos. 2001. Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52(7):548–557.

Sine Zambach and Tine Lassen. 2010. A lexical framework for semantic annotation of positive and negative regulation relations in biomedical pathways. In Nigel Collier, Udo Hahn, Dietrich R. Schuhmann, Fabio Rinaldi, Sampo Pyysalo, Nigel Collier, Udo Hahn, Dietrich R. Schuhmann, Fabio Rinaldi, and Sampo Pyysalo, editors, *Semantic Mining in Biomedicine*, volume 714 of *CEUR Workshop Proceedings*. CEUR-WS.org.