Marine Variable Linker: Exploring Relations between Changing Variables in Marine Science Literature

Erwin Marsi[1], **Pinar Øzturk**[1], **Murat V. Ardelan**[2]
Department of Computer Science[1], Department of Chemistry[2]
Norwegian University of Science and Technology
{emarsi,pinar,murat.v.ardelan}@ntnu.no

Abstract

We report on a demonstration system for text mining of literature in marine science and related disciplines. It automatically extracts variables (e.g. *CO2*) involved in events of change/increase/decrease (e.g *increasing CO2*), as well as co-occurrence and causal relations among these events (e.g. *increasing CO2 causes a decrease in pH in seawater*), resulting in a big knowledge graph. A web-based graphical user interface targeted at marine scientists facilitates searching, browsing and visualising events and their relations in an interactive way.

1 Introduction

Progress in science relies significantly on the premise that – in addition to other methods for gaining knowledge such as experiments and modelling – new knowledge can be inferred by combining existing knowledge found in the literature. Unfortunately such knowledge often remains undiscovered because individual researchers can realistically only read a relatively small part of the literature, typically mostly in the narrow field of their own expertise (Swanson, 1986). Therefore we need software to help researchers managing the ever growing scientific literature and quickly fulfil their specific information needs. Even more so for "big problems" in science, such as climate change, which require a system-level, cross-disciplinary approach.

Text mining of scientific literature has been pioneered in biomedicine and is now finding its way to other disciplines, notably in the humanities and social sciences, holding the promise for knowledge discovery from large text collections. Still, multidisciplinary fields such as marine science, climate science and environmental science remain mostly unexplored. Due to significant differences between the conceptual frameworks of biomedicine and other disciplines, simply "porting" the biomedical text mining infrastructure to another domain will not suffice. Moreover, the type of questions to be asked and the answers expected from text mining may be quite different.

Theories and models in marine science typically involve changing variables and their complex interactions, which includes correlations, causal relations and chains of positive/negative feedback loops, where multicausal events are common. Many marine scientists are thus interested in finding evidence – or counter-evidence – in the literature for events of change and their relations. Here we report on an end-user system, resulting from our ongoing work to automatically extract, relate, query and visualise events of change and their direction of variation.

Our text mining efforts in the marine science domain are guided by a basic conceptual model described in (Marsi et al., 2014). The system presented here covers a subset of this model, namely, change events, variables and causal relations. A *change* is an event in which the value of a variable is changing, but the direction of change is unspecified. There are two specific subtypes of a change event: an *increase* in which direction of change is positive and a *decrease* in which the direction of change is negative. A *variable* is something mentioned in the text that is changing its value. This is a very broad definition that covers much more than traditional entities (e.g. person, disease or protein) and includes long and complex expressions. A *cause* is relation that holds between a pair of events in which a change in one variable causes a change in another variable. An example of a sentence annotated according to this conceptual model is shown in Figure 1.

Proceedings of the EACL 2017 Software Demonstrations, Valencia, Spain, April 3-7 2017, pages 91–94

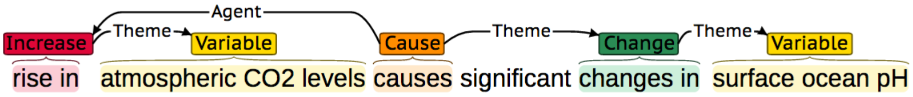


Figure 1: Example of text annotation according to conceptual model

2 Text mining system

Our text mining system for marine science literature is called *Megamouth*, inspired by the filter-feeder sharks which filter plankton out of the water. Its overall task is to turn unstructured data (text) into structured data (graph database) adhering to the conceptual model (discarding all other information) through a process of information extraction. The process is essentially a pipeline of processing steps, as briefly described below.

Step 1: Document retrieval involves crawling the websites for a predefined set of journals and extracting the text segments of interest from the HTML code, which includes title, authors, abstract, references, etc. Marine-related articles are selected through a combination of term matching with a manually compiled list of key words and a LDA topic model.

Step 2: Linguistic analysis consists of tokenisation, sentence splitting, lemmatisation, POS tagging and constituency parsing using the Stanford CoreNLP tools (Manning et al., 2014). It provides essential information required in subsequent processing such as variable extraction by pattern matching against syntactic parse trees.

Step 3: Variable and event extraction is performed simultaneously through tree pattern matching, where manually written patterns are matched against lemmatised constituency trees of sentences to extract events (increase/decrease/change) and their variables. It depends on two tools that are part of CoreNLP: Tregex is a library for matching patterns in trees based on tree relationships and regular expression matches on nodes; Tsurgeon is a closely related library for transforming trees through sequences of tree operations. For more details, see (Marsi and Øzturk, 2016; Marsi and Öztürk, 2015).

Step 4: Generalisation of variables addresses variables that are very long and complex and therefore unlikely to occur more than once. These are generalised (abstracted) by removing non-essential words and/or splitting them into atomic variables. For example, the variable *the annual, Milankovitch and continuum temperature* is split into three parts, one of which is *annual temperature*, which is ultimately itself generalised to *temperature*. This is accomplished through progressive pruning of a variable's syntactic tree, using a combination or tree pattern matching and tree operations.

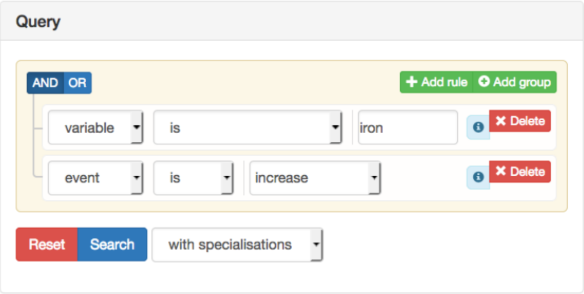
Step 5: Relation extraction again uses tree-pattern matching with hand-written patterns to extract causal relations between pairs of events, identifying their cause and relation roles.

Step 6: Conversion to graph All extracted variables, events and relations are subsequently converted to a single huge property graph, which is stored and indexed in a Neo4j graph database[1] (Community Edition) to facilitate fast search and retrieval. It contains nodes for variables, generalised variables, event instances, event relations, sentences and articles. It has edges between, e.g., a variable and its generalisations. Properties on nodes/edges hold information like a sentence's number and character string on sentence nodes, or the character offsets for event instances.

Step 7: Graph post-processing enriches the initial graph in a number of ways using the Cypher graph query language. Event instance nodes are aggregated in event type nodes. Likewise, causal relation instances are aggregated in causal relations types between event types. Furthermore, co-occurrence counts for event pairs occurring in the same sentence are computed and added as co-occurrence relations between their respective event type nodes. Post-processing also includes addition of metadata and citation information, obtained through the Crossref metadata API, to articles nodes in the graph.

The final output is a big knowledge graph (millions of nodes) containing all information extracted from the input text. The graph can be searched in many different ways, depending on interest, using the Cypher graph query language. One possibility is searching for chains of causal relations. The user interface described in the next section offers a more user-friendly way of searching for a certain type of patterns, namely, relations between changing variables.

[1] `https://neo4j.com/`



variable | is | iron

event | is | increase

with specialisations

Figure 2: Example of event query composition

3 User interface

Although graph search queries can be written by hand, it takes time, effort and a considerable amount of expertise. In addition, large tables are difficult to read and navigate, lacking an easy way to browse the results, e.g., to look up the source sentences and articles for extracted events. Moreover, users need to have a local installation of all required software and data. The Marine Variable Linker (MVL) is intended to solve these problems. Its main function is to enable non-expert users (marine scientists) to easily search the graph database in an interactive way and to present search results in a browsable and visual way. It is a web application that runs on any modern platform with a browser (Linux, Mac OS, Windows, Android, iOS, etc). It is a graphical user interface, which relies on familiar input components such as buttons and selection lists to compose queries, and uses interactive tables and graphs to present search results. Hyperlinks are used for navigation and to connect related information, e.g. the webpage of the source journal article.

Figure 2 shows an example of a search query for events consisting of two search rules: (1) the variable equals *iron* and (2) the event type is *increase*. In addition, the search is *with specialisations*, which means that it includes variables that can be generalised to *iron*, such as *particulate iron* or *iron in clam mactra lilacea*. More rules can be added to narrow down, or widen, event search. Clicking the search button will bring up a new table for matching event types, showing the instance counts, predicates and variables (not shown here). Clicking on any row in this event types table will bring up the corresponding event instances table, which shows all the actual mentions of this event in journal articles. Each row in the instance table shows a sentence, year of publication and source.

Once events are defined, one can search for relations between these events, where queries can be composed in a similar fashion as for events. The first kind of relation is *cooccurs*, which means that two events co-occur in the same sentence. When two events are frequently found together in a single sentence, they tend to be associated in some way, possibly by correlation. The second kind of relation is *causes*, which means that two events in a sentence are causally related, where one event is the cause and other the effect. Causality must be explicitly described in the sentence, for example, by words such as *causes*, *therefore*, *leads to*, etc.

Relation search results are presented in two ways. The relation types table contains all pairs of event types, specifying their relation, event predicates, event variables and counts. Figure 3 provides an example for an open-ended search query where the cause is an event with variable *iron* (including its specialisations, direction of change unspecified), whereas the effect is left open (i.e., can be any event). The corresponding relation graph is shown in Figure 4. The nodes are event types with red triangles for increases, blue triangles for decreases and green diamonds for changes.

Clicking on a row in the table or a node in the graph brings up a corresponding instances table (cf. bottom of Figure 3), showing sentences, years and citations of articles containing the given relation. The events are marked in colour: red for increasing, blue for decreasing and green for changing. Hovering the mouse over the document icon will show a citation for the source article, whereas clicking it will open the article's web page containing the sentence in a new window.

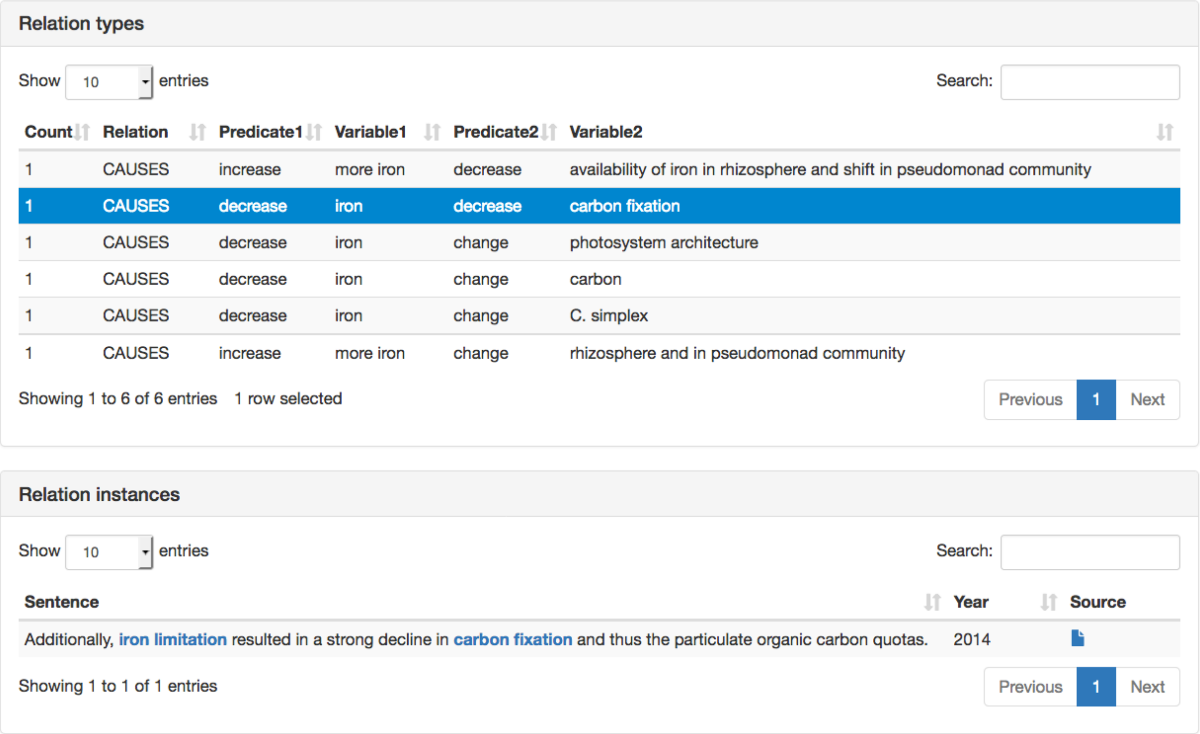
A demo of an MVL instance indexing 75,221 marine-related abstracts from over 30 journals is currently freely accessible on the web.[2] Source code for the text mining system and the graphical user interface is freely available.[3]

4 Discussion and Future work

Our system still makes many errors. Variables and events are sometimes incorrectly extracted (e.g. variable *more iron* in Figure 3 ought to be just *iron*), often due to syntactic parsing errors, and many are missed altogether (e.g. the decline in *the particulate organic carbon quotas* in the same Figure), because events can be expressed in so

[2] `http://baleen.idi.ntnu.no/demos/megamouth-abs/`

[3] `https://github.com/OC-NTNU`



Relation types

Show 10 entries

Search:

Count	Relation	Predicate1	Variable1	Predicate2	Variable2
1	CAUSES	increase	more iron	decrease	availability of iron in rhizosphere and shift in pseudomonad community
1	CAUSES	decrease	iron	decrease	carbon fixation
1	CAUSES	decrease	iron	change	photosystem architecture
1	CAUSES	decrease	iron	change	carbon
1	CAUSES	decrease	iron	change	C. simplex
1	CAUSES	increase	more iron	change	rhizosphere and in pseudomonad community

Showing 1 to 6 of 6 entries 1 row selected

Previous 1 Next

Relation instances

Show 10 entries

Search:

Sentence	Year	Source
Additionally, **iron limitation** resulted in a strong decline in **carbon fixation** and thus the particulate organic carbon quotas.	2014	🗎

Showing 1 to 1 of 1 entries

Previous 1 Next

Figure 3: Example of search results for causal relations types (top) and a selected instance (bottom)

many different and complex ways. Yet we feel that even with a fair amount of noise, the current proof-of-concept already offers practical merit in battling the literature deluge. We will continue to work on improving recall and precision, as well as usability aspects of the interface. One aspect under active development is the integration of new and better algorithms for causal relation extraction based on machine learning from manually annotated data. The knowledge graph can be searched efficiently using Cypher queries, which opens up many other interesting opportunities for knowledge discovery. We hope our system will attract interest from marine and climate scientists, raising awareness of the potential of text mining, as progress will crucially depend on building a community similar to that in biomedical text mining.

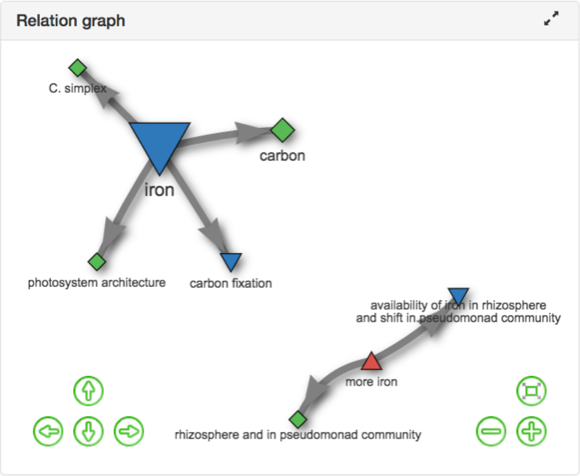

Relation graph
C. simplex
carbon
iron
photosystem architecture
carbon fixation
availability of iron in rhizosphere and shift in pseudomonad community
more iron
rhizosphere and in pseudomonad community

Figure 4: Example of relation graph

Acknowledgments

Financial aid from the European Commission (OCEAN-CERTAIN, FP7-ENV-2013-6.1-1; no:603773) is gratefully acknowledged.

References

C.D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S.J. Bethard, and D. McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL System Demonstrations*, pages 55–60.

E. Marsi and P. Öztürk. 2015. Extraction and generalisation of variables from scientific publications. In *EMNLP*, pages 505–511, Lisbon, Portugal.

E. Marsi and P. Øzturk. 2016. Text mining of related events from natural science literature. In *Workshop on Semantics, Analytics, Visualisation: Enhancing Scholarly Data (SAVE-SD)*, Montreal, Canada.

E. Marsi, P. Öztürk, E. Aamot, G. Sizov, and M.V. Ardelan. 2014. Towards text mining in climate science: Extraction of quantitative variables and their relations. In *Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*, Reykjavik, Iceland.

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