Text mining of related events from natural science literature

Erwin Marsi and Pinar Øzturk

Norwegian University of Science and Technology
Department of Computer and Information Science
Trondheim, Norway
{emarsi,pinar}@idi.ntnu.no

Abstract. We present an approach to text mining in areas where the entities of interest can not be defined in advance. Our system is aimed at finding related events in natural science literature, in particular, changing/increasing/decreasing variables in Marine science publications. It enables semantic search for events by abstracting from morphological, lexical-semantic and syntactic variations. In addition, generalisation of variables through syntactic pruning helps finding similar variables. Relations between events are induced from co-occurrence frequencies. Extracted information is stored in a property graph database and accessed using the Cypher query language. A user interface presents events as a graph to visualise their type, frequency and relation strength, in combination with their textual sources.

Keywords: text mining, natural language processing, information extraction, visualisation, knowledge discovery

1 Introduction

Text mining of scientific literature originates from efforts to cope with the ever growing flood of publications in biomedicine [1]. Consequently the resulting approaches, methods, resources and applications are rooted in the paradigm of biomedical research and its conceptual framework [2]. Text mining is now finding its way to other scientific disciplines, promising support for knowledge discovery from large text collections. Our own research targets text mining in marine science. As text mining efforts in this area are extremely rare [7, 6, 5], it is not surprising that a corresponding infrastructure is mostly lacking. Moreover, we found that due to significant differences between the conceptual frameworks of biomedicine and marine science, simply "porting" the biomedical text mining infrastructure will not suffice. One major difference is that the biomedical entities of interest are relatively well defined – genes, proteins, organisms, species, drugs, diseases, etc. – and typically expressed as proper nouns. In contrast, defining the entities of interest in marine science turns out to be much harder. Not only does it seem to be more open-ended in nature, the "entities" themselves tend to be complex and expressed as noun phrases containing multiple modifiers, giving

rise to examples like *timing and magnitude of surface temperature evolution in the Southern Hemisphere in deglacial proxy records.*

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. Many marine scientists are interested in finding evidence – or counter-evidence – in the literature for events of change and their relations. Given the difficulties with defining entities, we focus on mining of these events, leaving entities underspecified for the time being, simply referring to them as *variables*. Here we describe ongoing work to automatically extract, relate, query and visualise events of change and their direction of variation: *increasing*, *decreasing* or just *changing* (i.e. direction not specified in the text).

2 Approach

Our system is essentially a pipeline involving a number of processing steps.

Information retrieval – The first step is collecting publications of interest – for our use case, Marine science articles concerning the biological pump and/or food webs. Our text material consists of abstracts from selected journals by Nature Publishing Group. Search terms obtained from domain experts were used to query Nature's OpenSearch API[1] for publications in selected journals, after 1997, retrieving records including title and abstract. The top-10k abstracts matching most search terms were selected for further processing with the Stanford CoreNLP tools [4], including tokenisation, sentence splitting, part-of-speech tagging, lemmatisation and syntactic parsing. Lemmatised parse trees were obtained by substituting terminals with their lemmas. The resulting new corpus contains 9,884 article abstracts, 29,565 sentences and approximately 626k tokens.

Information extraction – The second step extracts change events and the variables they pertain to. Tree pattern matching is applied to lemmatised syntax trees using the Tregex engine [3], which provides a compact language for writing regular expressions over trees. Seven hand-written pattern templates were instantiated with lexical instances from manually created lists of verbs and nouns expressing change, yielding 320 tree matching patterns. The total number of matched variables in the corpus is 22,784: 9,673 for change, 7,827 for increase and 5,289 for decrease. For more details, see [5].

Generalisation of variables – Since many of the extracted variables are long and complex expressions, their frequency is low. The most frequent variables are generic terms (*climate* 1350, *temperature* 165, *global climate* 86), but over 66% is unique. This evidently impedes the discovery of relations among events. As a partial solution to this problem, variables are generalised by progressive pruning of syntax trees using a set of tree transformation operations. This effectively produces more abstract variants of variables. For example, the variable *the annual, Milankovitch and continuum temperature* is split into three variables –

[1] http://www.nature.com/developers/documentation/api-references/opensearch-api

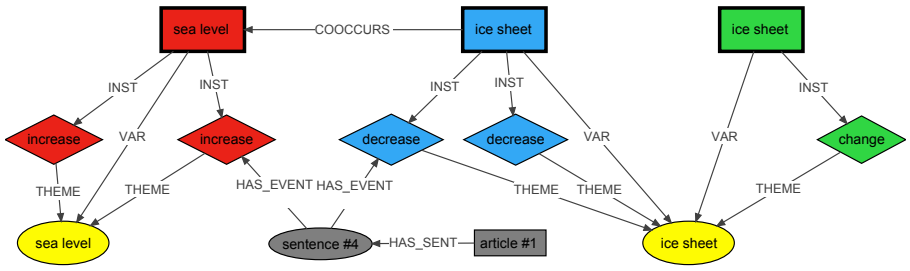

Fig. 1. Event model in the graph database

annual temperature, *Milankovitch temperature* and *continuum temperature* – all which are ultimately reduced to just *temperature*. Tree transformations are implemented using Tsurgeon [3]: Tregex patterns match the syntactic structures of interest, whereas an associated Tsurgeon operation deletes selected nodes (cf.[5]) Generalisation resulted in 102,625 variables, which is 4.5 times the number of originally extracted variables.

Graph creation – The extracted events are stored in a property graph database as nodes, directed edges and associated properties. Figure 1 shows a small partial sample of how events are modelled. The diamond-shaped nodes represent events, with red for an INCREASE, blue for a DECREASE and green for a CHANGE events. An event pertains to a unique VARIABLE type (yellow nodes) as indicated by its THEME edge. Each event also occurs in a SENTENCE through a HAS-EVENT edge, which in turn is linked to an ARTICLE via a HAS-SENT edge. Aggregated nodes join all event nodes with the same type and variable. For example, the square blue node labelled "ice sheet" joins all event instances where the variable "ice sheet" is decreasing. Likewise, the square red node labelled "sea level" joins all sea level decrease events. Finally, aggregated nodes can be connected by COOCCURS edges whenever two events co-occur in a single sentence. For example, a decrease of "ice sheet" co-occurs with an increase of "sea level" in sentence number 4. In addition, nodes and edges have properties which hold important information. For instance, SENTENCE nodes hold the sentence string, event nodes hold the character offsets for their variable string and COOCCURS edges hold the frequency of co-occurrence. The Neo4j graph database[2] (community edition) is used for storing and accessing the graph. The powerful Cypher query language – akin to SQL for relational databases – makes it relatively easy to search for relations between events, for example, to find the shortest path between an increase in A and a decrease in B over any number of co-occurrence links.

User interface These search capabilities are partly exposed to end users through a web application with a graphical user interface. Users can search for single events, relations between pairs of events or even triples of related events (i.e. indirect relations). Each event can optionally be restricted by type (increase, decrease or change) and by variable involved. The type of relation between events

[2] http://neo4j.com/

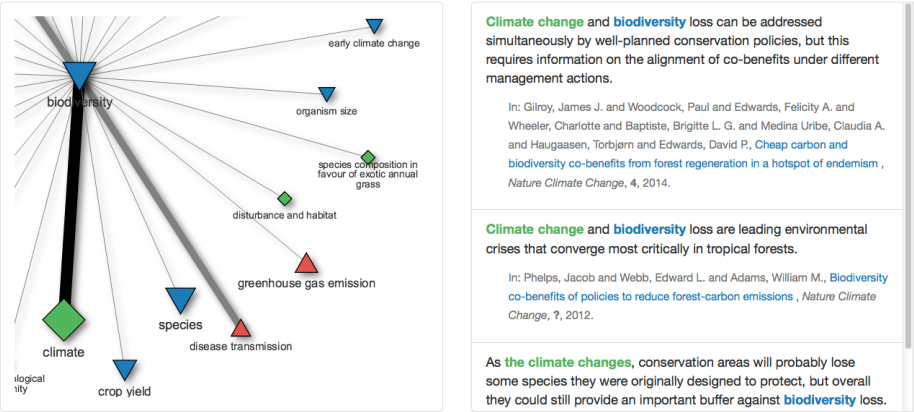

Fig. 2. Visualisation of event relations

is currently limited to co-occurrence, but will be extended to causal relations and correlations in the near future. Figure 1 shows part of the output for any events related to a *decrease in biodiversity*. The left pane shows a graph where the nodes are events – red for increase, blue for decrease and green for change – labelled with their variable.[3] The node size corresponds to an event's overall frequency, whereas edge weight denotes co-occurrence frequency. The graph can be moved/resized and otherwise filtered and formatted according to a user's need. Selecting an edge (black edge in Figure 1) will list the corresponding source sentences in the right pane, with highlighted variables and links to original article web pages.

3 Discussion

We have presented an approach to text mining from natural science literature which is aimed at finding related events. It provides semantic search for events in the sense that it abstracts from morphological variations (e.g. singular/plural), lexical-semantic variations (e.g. an increase can be expressed by *rise, enhance, boost*, etc.), syntactic variations (e.g. *X increases, something increases X, increasing X, X is increasing, an X increase, increase in X*, etc.). In addition, generalisation of variables through syntactic pruning helps finding similar variables: for example, both *the annual, Milankovitch and continuum temperature variability* and *annual temperature between 1958 and 2010* are progressively generalised to *annual temperature*, revealing their similarity at a more abstract level. Whereas a more elaborate description of the information retrieval, extraction and generalisation steps was presented in [5], novel contributions here include the graph model and the user interface. Events, variables and other information are stored in a property graph database and can thus be easily accessed, traversed or modified using the Cypher query language. A user interface presents

[3] Graph rendering with vis.js Javascript library: http://visjs.org/.

events as a graph to visualise their type, frequency and relation strength, also providing links their textual sources. We believe the approach is general and applicable to other areas where the entities of interest can not be defined in advance (with minor adaptations of patterns and lexical items).

The current implementation is a proof of concept, but produces a fair amount of noise. Analysis suggests that most problems originate from syntactic parsing errors (in particular coordination and prepositional phrase attachment). As a result, patterns may either fail to match or match unintentionally, yielding incomplete or incoherent variables. Pruning variables is beneficial but insufficient and should be supplemented with other methods. For instance, linking named entities like species, chemicals or geographical locations to unique concepts in appropriate ontologies/taxonomies would allow for generalisations such as *iron* is a *metal* or a *diatom* is a *plankton*. Likewise co-occurrence frequency is a weak signal and part of our ongoing work is therefore to extract causal relations and correlations between events using both pattern matching and machine learning methods. Ultimately events obtained from different publications can be chained together, often with the help of domain knowledge, in order to generate new hypotheses, as pioneered in the work on literature-based knowledge discovery [8]. We will release the source code of a more mature version of our software, as well as various data sets of extracted events, in the near future.

Acknowledgements – Financial aid from the European Commission (OCEAN-CERTAIN, FP7-ENV-2013-6.1-1; no: 603773) is gratefully acknowledged. We thank Murat Van Ardelan for sharing his knowledge of Marine science.

References

1. Ananiadou, S., Mcnaught, J.: Text Mining for Biology And Biomedicine. Artech House, Inc., Norwood, MA, USA (2005)
2. Cohen, K.B., Hunter, L.: Getting Started in Text Mining. PLoS Computational Biology 4(1), e20+ (2008)
3. Levy, R., Andrew, G.: Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In: Proc. of ELREC. pp. 2231–2234 (2006)
4. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proc. of ACL. pp. 55–60 (2014)
5. Marsi, E., Öztürk, P.: Extraction and generalisation of variables from scientific publications. In: Proc. of EMNLP. pp. 505–511. Lisbon, Portugal (2015)
6. Marsi, E., Öztürk, P., Aamot, E., Sizov, G., Ardelan, M.V.: Towards text mining in climate science: Extraction of quantitative variables and their relations. In: Proc. of Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing. Reykjavik, Iceland (2014)
7. Radom, M., Rybarczyk, A., Kottmann, R., Formanowicz, P., Szachniuk, M., Glöckner, F.O., Rebholz-Schuhmann, D., Błażewicz, J.: Poseidon: An information retrieval and extraction system for metagenomic marine science. Ecological Informatics 12, 10–15 (2012)
8. Swanson, D.R.: Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in biology and medicine 30(1), 7–18 (1986)