Natural Language Processing and Machine Learning: Applying Advances in BioMedicine to Earth Science

Ruth Duerr, Skatje Myers, Martha Palmer, Chris Jenkins, Anne Thessen, James Martin

An NSF-Funded project at INSTAAR and CLEAR at the University of Colorado Boulder, the Smithsonian Institute, and Ronin Institute for Independent Scholarship

What is ClearEarth ?

ClearEarth is a collaborative project aimed at bringing semantic technologies from the biomedical field into the earth-surface earth, ice and life sciences.

Why Biomedicine ?

Biomedicine is very advanced in building large, deep semantic resources using Natural Language Processing (NLP) and Machine Learning (ML) techniques.

Goals

Overall

Enable use of advanced industry and research software within the geo-, bio- and cryospheric sciences for operations such as query and reasoning.

Year 1

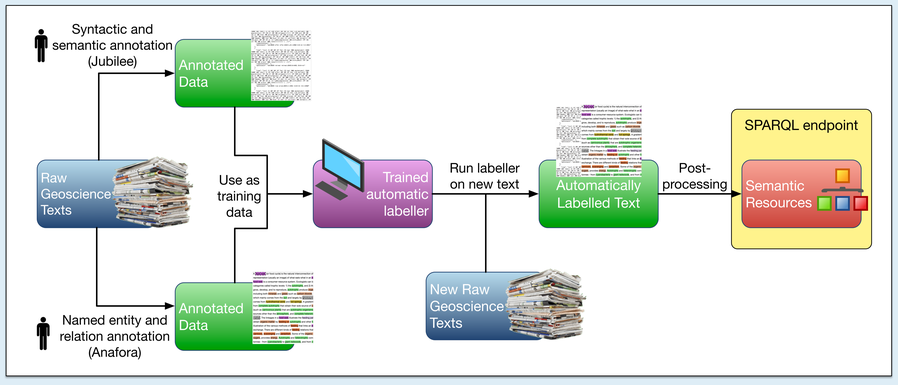
- Develop efficient and accurate methods of annotating earth-ice-biology texts with syntactic and semantic content.
- Investigate pathways for successful adoption of the NLP and ML methods from the biomedical semantic technologies.

Early Lessons

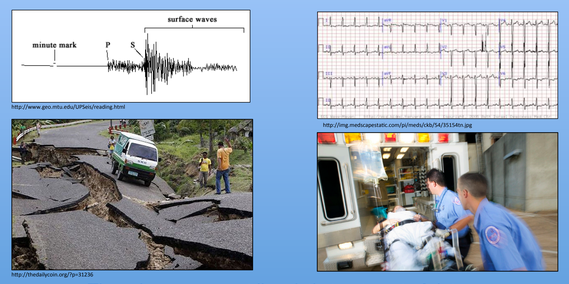
- Obtaining high quality, internally consistent annotation which is useful to machine learning algorithms is non-trivial
- Thought structures and concepts used by domain scientists are different and messier than those required by machines
- Interdisciplinary collaboration between highly disparate groups is challenging: terminologies and priorities
- Genres of text are distinguished: for example, professional publication, popular article, newscast, twitter, field notes. We find all these in the geo-bio-cryo fields, just as much in biomedicine.

Methods

The most successful current approach for automatically extracting semantic content from text is by using supervised machine learning.

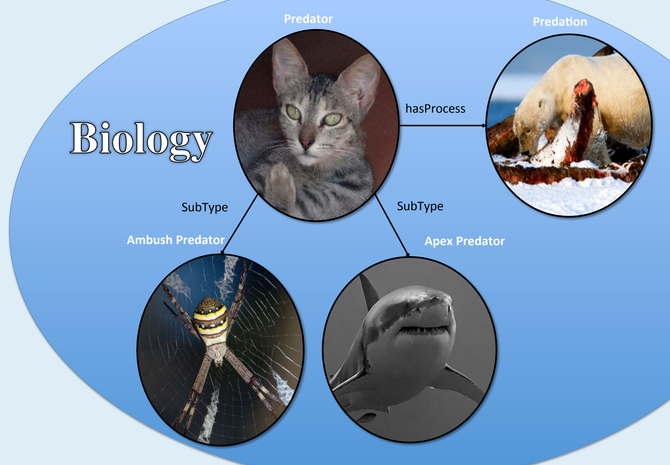


Geology



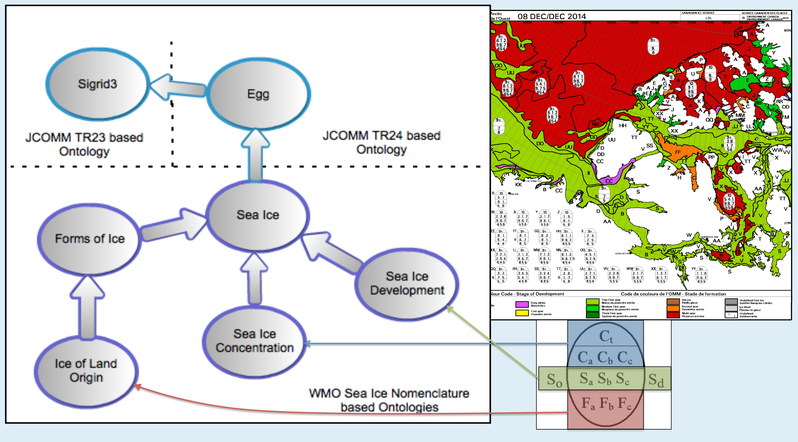
There is a congruent linguistic treatment of data on 'Events' between biomedicine and the natural sciences. We exploit that for the annotation process, and later in the NLP-ML software.

Biology



Validation

Verify that these approaches can re-create hand built ontologies developed where both methods use the same source documents. Using ontologies developed through the Semantic Sea Ice Interoperability Initiative funded Award ACI 0956010 INTEROP (as depicted below).

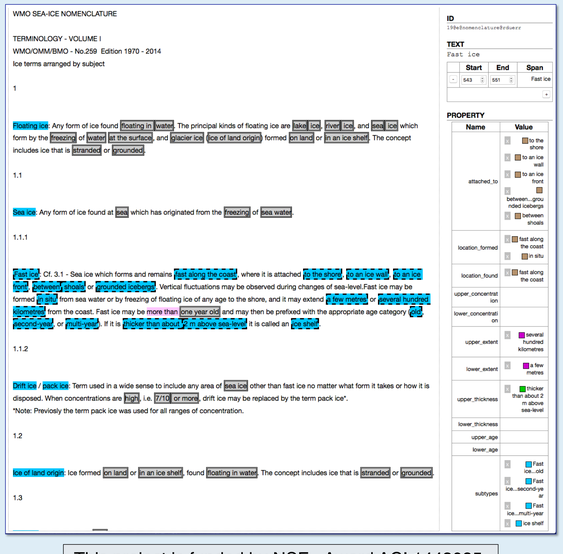


Cryospheric Science

Beyond the WMO Sea Ice Nomenclature, other cryospheric glossaries are being gathered as input corpora for this project. For example, glossaries of permafrost[1], snow terms[2], glacier[3] and glacier mass balance[4] terminology, and general cryospheric glossary[5].

We welcome pointers to other glossaries!!!!

Anafora Annotation



The extreme importance of the annotation (here, Anafora) stage must be emphasized. The Machine Learning (ML) software ported from BioMedicine requires sufficiently large (large !) annotated corpus. Noisy entries work against it, so annotation must be careful. But once that threshold is reached, virtually all texts in the domain are treatable by the ML-NLP methods.

[1] van Everdingen, Robert, ed. 1998 revised May 2005. Multi-language glossary of permafrost and related ground-ice terms. Boulder, CO: National Snow and Ice Data Center. Available at https://nsidc.org/fgdc/glossary/
[2] C. Fierz, R. L. Armstrong, Y. Durand, P. Etchevers, E. Greene, D. M. McClung, K. Nishimura, P. K. Satyawali and S. A. Sokratov, 2008, IACS International Classification for Seasonal Snow on the Ground. Available at http://www.aineva.it/previsori/Classificazione%20Internazionale/20080720_iacs_classif2008.pdf
[3] B. Molnia, 2004, Glossary of Glacier Terminology: A glossary providing the vocabulary necessary to understand the modern glacier environment, Open-File Report 2004-1216, USGS, available at http://pubs.usgs.gov/of/2004/1216/text.html
[4] Cogley, J.G., R. Hock, L.A. Rasmussen, A.A. Arendt, A. Bauder, R.J. Braithwaite, P. Jansson, G. Kaser, M. Möller, L. Nicholson and M. Zemp, 2011, Glossary of Glacier Mass Balance and Related Terms, IHP-VII Technical Documents in Hydrology No. 86, IACS Contribution No. 2, UNESCO-IHP, Paris. Available at http://unesdoc.unesco.org/images/0019/001925/192525E.pdf
[5] Cryosphere glossary, NSIDC. Available at http://nsidc.org/cryosphere/glossary

This project is funded by NSF - Award ACI 1443085.