C1F21 DIBBS: Porting Practical Natural Language Processing (NLP) and Machine Learning (ML)
Semantics from Biomedicine to the Earth, Ice and Life Sciences
AWARD #ACI 1443085

Chris Jenkins, Jim Martin, Martha Palmer, Skatje Myers, Ruth Duerr, Sarah Ramdeen,
Anne Thessen, Jenette Preciado; University of Colorado & Ronin Institute.

CHALLENGES AND FUTURE DIRECTIONS

Annotation, Schema, Genres – The heavy load of training the Machine Learning / Natural Language Processing (ML/NLP) has to be streamlined. The project is making steps in this direction, in a sense trailblazing for general earth sciences using the 3 domains geology, cryology, ecology. The most promising step will be machine-assisted pre-annotation, while also improving the productivity of manual inputs.

Higher adjudication scores need to be obtained more quickly. A closer interfacing between the annotation work and scores from early processings will help. This is a matter of implementing smart workflows taking advantage, in a sequence, of man and machine.

Improving the ML/NLP processes - An example is with deep learning-based word embeddings**.** These techniques have the advantage of being able to leverage large amounts of unannotated data, an advantage that will be critical as we move on to exploring semi-supervised approaches to improving the performance of the baseline systems. However, at the moment we have not collected large enough sets of domain specific documents for the word embeddings to be effective.

Information extraction – The end-goal of the project. As is often pointed out, the amount of free-text information in each science is huge. Furthermore, that data is qualitatively different from numeric data types. An example would be information on the traits of organisms, even extending to their behaviors. One consideration is how the science community want the information delivered – database, triple store, ontology. But a bigger challenge is designing and implementing suitable, effective QA/QC filters before releasing the information products.

Rollout of techniques to the wider domain - Develop a well-organized process for extending ML/NLP methods to other earth-science sub-domains covering materials, processes, organisms, structures. We need to disseminate the benefits of what we have learned in this project (e.g., on workflows) in a variety of steps including conferences, papers, hackathons. Part of the challenge is bridging the divide between domainal teams (e.g., geologists) and the computational scientists. Having startlingly interdisciplinary projects like ClearEarth is important, also involving students early-on in such projects.