Automated Trait Extraction using ClearEarth, a Natural Language Processing System for Text Mining in Natural Sciences

Anne E Thessen†§, Jenette Preciado⡁, Payoj Jain⡁, James H Martin⡁, Martha Palmer⡁, Riyaz Bhat⡁

† The Ronin Institute for Independent Scholarship, Monclair, NJ, United States of America
‡ The Data Detektiv, Waltham, MA, United States of America
¶ University of Colorado Boulder, Boulder, CO, United States of America

Corresponding author: Anne E Thessen (annethessen@gmail.com)

Received: 22 Apr 2018 | Published: 22 May 2018


Abstract

The cTAKES package (using the ClearTK Natural Language Processing toolkit Bethard et al. 2014, http://cleartk.github.io/cleartk/) has been successfully used to automatically read clinical notes in the medical field (Albright et al. 2013, Styler et al. 2014). It is used on a daily basis to automatically process clinical notes and extract relevant information by dozens of medical institutions. ClearEarth is a collaborative project that brings together computational linguistics and domain scientists to port Natural Language Processing (NLP) modules trained on the same types of linguistic annotation to the fields of geology, cryology, and ecology. The goal for ClearEarth in the ecology domain is the extraction of ecologically-relevant terms, including eco-phenotypic traits from text and the assignment of those traits to taxa. Four annotators used Anafora (an annotation software; https://github.com/weitechen/anafora) to mark seven entity types (biotic, aggregate, abiotic, locality, quality, unit, value) and six reciprocal property types (synonym of/has synonym, part of/has part, subtype/supertype) in 133 documents from primarily Encyclopedia of Life (EOL) and Wikipedia according to project guidelines (https://github.com/ClearEarthProject/AnnotationGuidelines). Inter-annotator agreement ranged from 43% to 90%. Performance of ClearEarth on identifying named entities in biology text overall was good (precision:...
The named entities with the best performance were organisms and their parts/products (biotic entities - precision: 72.09%; recall: 54.17%) and systems and environments (aggregate entities - precision: 79.23%; recall: 75.34%). Terms and their relationships extracted by ClearEarth can be embedded in the new ecocore ontology after vetting ([http://www.obofoundry.org/ontology/ecocore.html](http://www.obofoundry.org/ontology/ecocore.html)). This project enables use of advanced industry and research software within natural sciences for downstream operations such as data discovery, assessment, and analysis. In addition, ClearEarth uses the NLP results to generate domain-specific ontologies and other semantic resources.

**Keywords**

text mining, NLP, phenotype, ecology, ontology

**Presenting author**

Anne E Thessen

**Presented at**

TDWG 2018

**Funding program**

National Science Foundation: Data Infrastructure Building Blocks

**Grant title**

DIBBS: Porting Practical NLP and ML Semantics from Biomedicine to the Earth, Ice and Life Sciences

**Hosting institution**

University of Colorado Boulder
References

