The ClearEarth Project: Preliminary Findings from Experiments in Applying the CLEARTK NLP Pipeline and Annotation Tools Developed for Biomedicine to the Earth Sciences

Ruth Duerr, Sarah Ramdeen, Anne Thessen, Chris Jenkins, Martha Palmer, and Skatje Myers

Computational Language and EducAtion Research
CLEAR

AGU Fall Meeting 2016 Poster #IN11B-1625

An NSF-Funded project (Award ACI 1443085) 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, ice and life sciences.

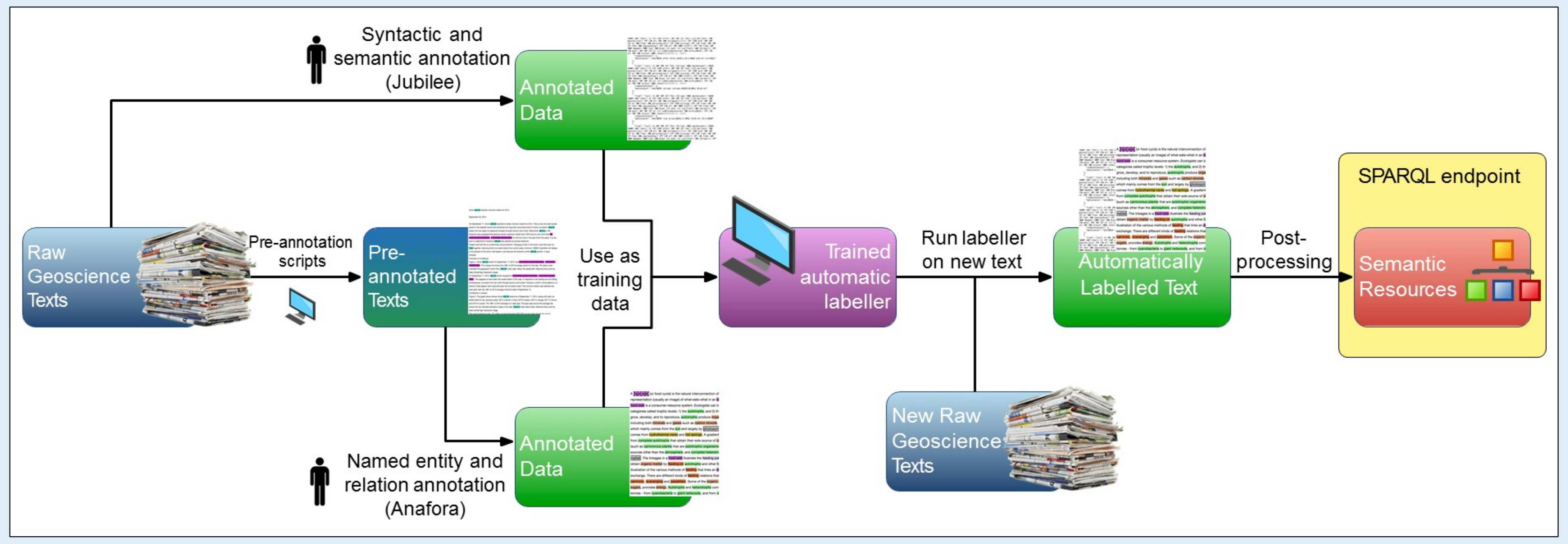
Goals

Overall: Enable use of advanced industry and research software within the geo-, bio- and cryo-spheric sciences for downstream operations such as query and reasoning.

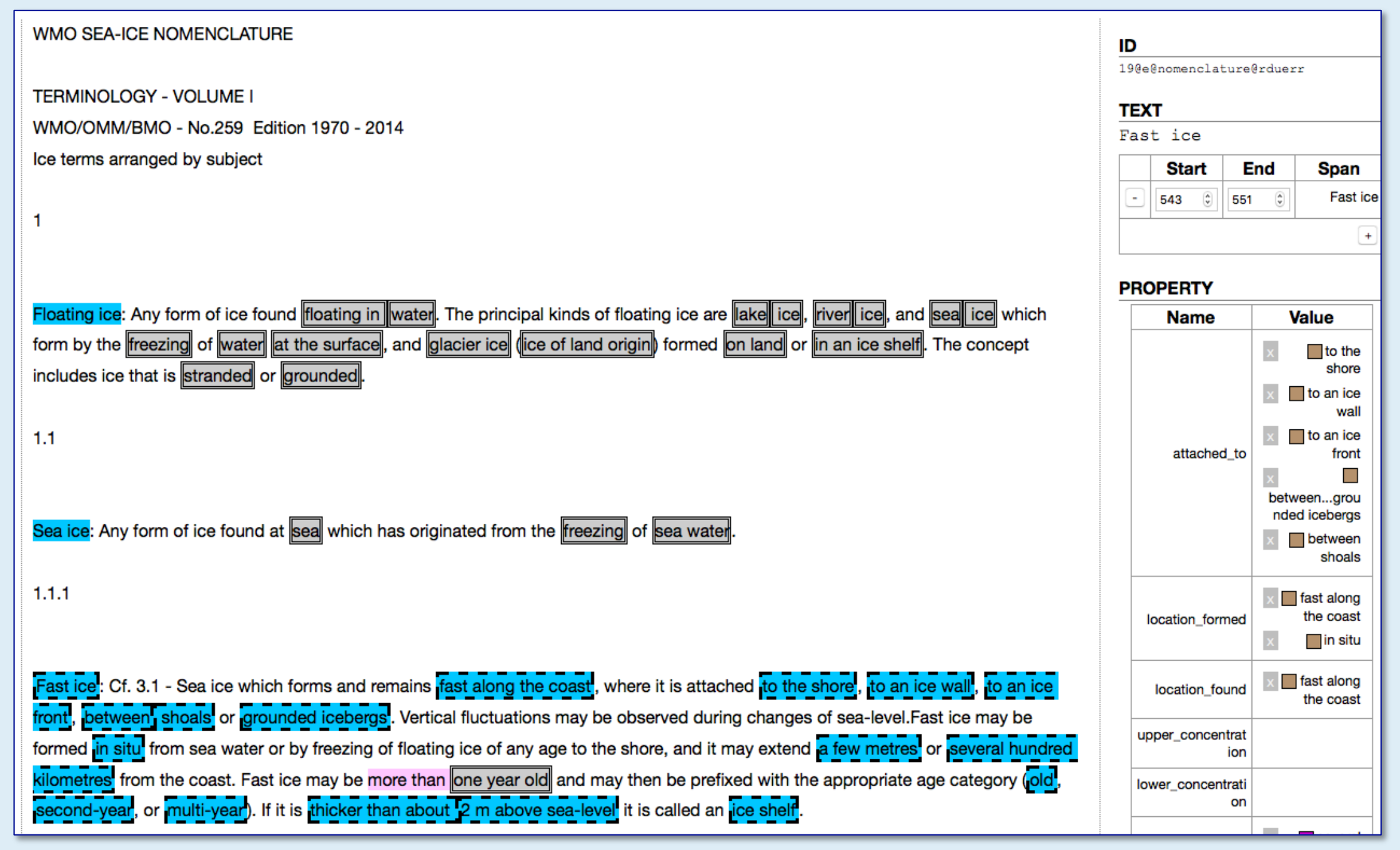
Year 2: 1) Develop efficient and accurate methods of annotating earth-ice-biology texts with syntactic and semantic content. 2) Investigate pathways for successful adoption of the NLP and ML methods from the biomedical semantic technologies.

Updated Methods

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



Anafora Annotation



WMO SEA-ICE NOMENCLATURE

TERMINOLOGY - VOLUME I

WMO/OMM/BMO - No.259 Edition 1970 - 2014

Ice terms arranged by subject

1

Floating ice: Any form of ice found floating in water. The principal kinds of floating ice are lake ice, river ice, and sea ice which form by the freezing of water at the surface, and glacier ice (ice of land origin) formed on land or in an ice shelf. The concept includes ice that is stranded or grounded.

1.1

Sea ice: Any form of ice found at sea which has originated from the freezing of sea water.

1.1.1

Fast ice: Cf. 3.1 - Sea ice which forms and remains fast along the coast, where it is attached to the shore, to an ice wall, to an ice front, between shoals or grounded icebergs. Vertical fluctuations may be observed during changes of sea-level.Fast ice may be formed in situ from sea water or by freezing of floating ice of any age to the shore, and it may extend a few metres or several hundred kilometres from the coast. Fast ice may be more than one year old and may then be prefixed with the appropriate age category (old, second-year, or multi-year). If it is thicker than about 2 m above sea-level it is called an ice shelf.

ID
19@e@nomenclature@rduerr

TEXT
Fast ice

	Start	End	Span
-	543	551	Fast ice

PROPERTY

Name	Value
attached_to	to the shore to an ice wall to an ice front between...grounded icebergs between shoals
location_formed	fast along the coast in situ
location_found	fast along the coast
upper_concentration	
lower_concentration	

The annotation (Anafora) stage is extremely important. The Machine Learning (ML) software requires sufficiently large annotated corpus. Once a threshold is reached, virtually all texts in the domain are treatable by the ML-NLP methods.

Outcomes

Collaboration with IPA

The International Permafrost Association is working with the ClearEarth project to characterize the contents of their global permafrost database. They hope that characterization of the descriptions of each entry will enable them to improve discoverability of permafrost data, while meeting criteria as specified by users.

Tasks:

- Associate standardized tags for Hydrology; Landform; Lithology; and Morphology fields.
- Based on the free text in these fields, determine and tag each record to improve the search process.
- The IPA community will vet these tags as they did manually, previously, for the vegetation category.

Collaboration with YAMZ

YAMZ is a crowdsourced metadata dictionary we are using to add glossary terms from the various domains represented in our project (see www.yamz.net).

- ClearEarth has scripts to convert the glossaries to a form which can be uploaded to YAMZ
- YAMZ assigns to each term a permanent URL that then can be used by ontology developers and domain scientists.
- Definitions from each glossary are tagged with a citation to their source.
- Where multiple definitions exist in different glossaries, a single term will be created with each definition tagged to its source.
- This allows communities to see the differences and begin to resolve them
- Next steps include asking the various communities to use these permalinks in their glossaries, to include them when developing domain specific semantic resources, and to begin to resolve differences between glossary definitions where they exist.
- Eventually one could imagine semantic markup in journal articles to the definitions wherever the terms are used.

Glossaries Being Addressed	YAMZ Tag
Glossary of Glacier Terminology	#USGSGlaciers
Glossary of Glacier Mass Balance and related terms	#IHPGlacierMassBalance
IACS International Classification for Seasonal Snow on the Ground	#IACSSnow
Permafrost Glossary	#IPAPermafrost
Cryosphere Glossary	#NSIDCCryosphere
Sea ice nomenclature	#WMOSeaIce
Biology Glossary	#ClearEarthBio

Ideas for Speeding up the Annotation Process

Hypothesis: Using software to pre-annotate concepts, even multi-word expressions, from domain glossaries will speed up the annotation process. This allows us to use terms the community has already agreed upon to pre-annotate text before getting to Anafora. At that point annotators verify the pre-annotations and add additional annotations beyond the glossary content.

Gold standard annotation guidelines

Rationale:

- The science community writes and thinks differently about domain specific texts than the general populace or even researchers in different domains.
- This limits the conceptual meanings that can be extracted from text using annotators who are not domain specialists
- Annotation guidelines are needed to teach non-domain specialist annotators what and how to annotate these texts

Outcome:

- 'Gold standard' annotation guidelines were created for the cryo and bio domains.
- Creating "gold standards" ensures that the guidelines can be implemented by annotators who are not domain specialists while also ensuring their scientific rigor in the domain.
- Our metric for determining the efficacy of the guidelines is when inter-annotator agreement reaches at least 75%.
- The Sea Ice Guidelines have recently been baselined and will be added to the https://github.com/ClearEarthProject website

Annotation Lessons Learned

Annotation spans

- When annotating text NLP experts tend to use either
  - minimum span or
  - maximum span
- Scientific concepts in glossaries are often phrases
- Worse yet, scientific texts often combine several such concepts into a single compound phrase where the ordering of the words may not be optimal
- ClearEarth agreed that if a phrase could be rephrased and partitioned into glossary concepts each component would be annotated

Example mention of ice terms in the phrase

"…of thin Arctic first-year drift ice in late summer"

- "…of thin Arctic first-year drift ice in late summer
- "…of thin Arctic first-year drift ice in late summer
- *Thin first-year ice is a term in the WMO sea ice nomenclature,*
- So our "ice" mention becomes
- *"…of thin Arctic first-year drift ice in late summer*
- *And "Arctic" is annotated as a location*
- *And "drift ice", another term in the WMO sea ice nomenclature, is annotated as an ice concentration*

Pushing Beyond the State of the Art

Dealing with complicated measurements

Example: Old ice is sea ice which has survived "at least one summer's melt"

The age of a type of sea ice is based on two factors:

1. the number of summer's melts that it has survived
2. whether or not it is in a "new cycle of growth."

State of the art: "one summer's melt" would be annotated as:

```
SeasonOfYear("summer",
Type=Summer,
Number=Number("one",
Value=1))
```

Which really translates to "every summer!!!" Not what is intended!

Pushing Beyond: Experimenting with JPL measurement extraction software

Dealing with complex chains of processes and subsidiary processes

Example: Shear [stress]CAUSE causes rocks to [[slide past each other]CAUSE]EFFECT resulting in an [earthquake]EVENT and strike-slip [faults]EFFECTS

State of the art & pushing beyond:

- Tailoring RED[1] guidelines to these domains.

Dealing with negation

Example: The base species in a food web are those **species without prey** …

State of the art and pushing beyond:

- No straightforward solution
- Currently adding a "lack behavior" property for only the most important cases – address in future work

[1] Tim O'Gorman; Kristin Wright-Bettner; Martha Palmer, Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation, Computing News Storylines Workshop, held with EMNLP 2017, Austin, TX, November 2017