Annotating Event Chains for Carbon Sequestration Literature

Heng Ji[a], Xiang Li[a], Angelo Lucia[b], Jianting Zhang[c]

[a]Computer Science Department, Queens College and Graduate Center, City University of New York
Flushing, NY 11367, USA
E-mail: hengji@cs.qc.cuny.edu

[b]Department of Chemical Engineering, University of Rhode Island
Kingston, RI 02881-0805, USA
E-mail: lucia@egr.uri.edu

[a]Computer Science Department, City College of New York and Graduate Center, City University of New York
New York, NY 10031, USA
E-mail: jzhang@cs.ccny.cuny.edu

Abstract

In this paper we present a project of annotating event chains for an important scientific domain – carbon sequestration. This domain aims to reduce carbon emissions and has been identified by the U.S. National Academy of Engineering (NAE) as a grand challenge problem for the 21st century. Given a collection of scientific literature, we identify a set of centroid experiments; and then link and order the observations and events centered around these experiments on temporal or causal chains. We describe the fundamental challenges on annotations and our general solutions to address them. We expect that our annotation efforts will produce significant advances in inter-operability through new information extraction techniques and permit scientists to build knowledge that will provide better understanding of important scientific challenges in this domain, share and re-use of diverse data sets and experimental results in a more efficient manner. In addition, the annotations of metadata and ontology for these literature will provide important support for data lifecycle activities.

1. Introduction

The domain of carbon sequestration has been identified by the U.S. National Academy of Engineering (NAE, 2008) as a grand challenge problem for the 21st century. Carbon emissions from anthropogenic activities (power plants, transportation, manufacturing, etc.) have been linked to global warming and climate change. Carbon sequestration (iron fertilization storage in geological formations, ocean sediments, in plants and soil, etc.) refers to storage of CO_2 from anthropogenic sources and is generally preceded by CO_2 capture and transportation. A lot of research is being conducted on the methods of carbon sequestration and their short and long term impacts on the natural carbon cycle (carbon exchange between the atmosphere, terrestrial biosphere, oceans, and sediments).

As a result, hundreds of new papers and data sets are published in the carbon sequestration domain on a daily basis. It has become impractical for scientists to manually track all these new results and observations, and mine the data sets to construct a knowledge base. For example, if we search the query of "carbon sequestration", the Google Scholar search engine returns 92,300 relevant papers. However, based on our survey each researcher in this area is only able to read 1-2 papers on average every day, and spend about one hour on each paper. Furthermore, most of these papers are not freely available and thus the scientists are lack of effective tools to automatically distill the abstracts and select informative papers. More importantly, it has been quite challenging for the scientists to have a comprehensive view of what other relevant experiments have been done before they design a new experiment. Given the long cycle of a scientific experiment in this domain, it's critical to design new methods for efficiently mining and linking relevant results.

The sentences in these literature are often complex (long, multi-concept), ambiguous, flexible, and subtle. Achieving really high performance for annotation requires that we go beyond traditional data mining techniques, and conduct linguistic annotations to assist deep understanding of broad topics. Information extraction (IE) techniques can partially address these needs. IE can identify the instances of specified types of names/entities, relations and events from semi-structured or unstructured texts; and create a database. Besides trigger words, the relations and events also include the participants and their modifiers (date, time, location, etc.). The objective of extracting information from natural language has provided the opportunity to explore potential applications in military, medical, financial and bio-medical areas. Recent advances in cross-document IE

(Ji et al., 2009) make it possible to extract more salient, accurate and concise event information, such as tracking a person's employment history or a company's merger and acquisition activities.

The overall goal of our project is to adapt cross-document IE techniques to the needs of the carbon sequestration field. IE can play a significant role by automatically generating an accurate summary of facts and predicting new results, and thus assist scientists in selecting relevant papers, validating the correlations among different processes, and decision making. In addition, the extracted results from various related experiments are linked as event chains. The annotations of metadata and ontology for these literature will provide important support for data lifecycle activities. For example, scientists can track changes of experimental conditions including temperatures and volumes over time. By analyzing knowledge embedded in these papers, we can unleash a much more powerful knowledge resource than simple keyword search.

2. Motivation of Corpus Annotation

Most available IE corpora consist of news articles. Annotations for the carbon sequestration domain have not been captured in earlier efforts. We have applied a typical IE system trained from news articles (Ji and Grishman, 2008) directly to 50 paper abstracts from the Energy, Sustainability and Climate Change 2010 Conference, and didn't observe any correct extraction results due to the large difference across these two domains. For example, for the following sentence, this news-IE system mistakenly identified "Adaptive Online Control" and "Cascading Blackout Mitigation" as organization names because they were capitalized, while failed to identify the important relation between "Allocation Model" and "right hand side":

> *We presented <ORG>Adaptive Online Control </ORG> for <ORG>Cascading Blackout Mitigation </ORG> and resource reservation for Allocation Model with randomness on the right hand side.*

Jiang and Zhai (2007) reported similar "domain over-fitting" problems. They found that the F-measure of a name tagger trained from the "fly" domain significantly degraded from 54.1% to 28.1% when it's applied to the "mouse" domain.

Effective domain adaptation techniques may partially solve this problem, however, traditional domain adaptation methods assume some common "pivot" features between the source domain and target domain (Blitzer et al., 2006) and require certain prior knowledge about the target domain. Unfortunately both of these two assumptions don't hold for the carbon sequestration domain. For example, we found that only 36 of the 1233 ACE (NIST, 2005) event trigger words between the news domain and the carbon sequestration domain overlap. Therefore we will aim to attempt the once-popular sublanguage analysis scheme (Grishman, 2001) based on word class discovery and pattern learning, and also incorporating advanced machine learning and domain adaptation methods.

As the first modest step toward this goal, we have started a project to annotate event chains for a corpus including thousands of carbon sequestration literature (section 4). Then we focus on describing the detailed annotation challenges and our general solutions (section 5).

3. A General Vision

Given a collection of scientific documents written in natural language, our general goal is to identify a set of centroid experiments; and then link and order the observations and events centered around these experiments on a causal or temporal chain. What might such event chains look like? For example, from the following document (Brewer et al., 1999) about "CO2 Geological Sequestration":

> *Field experiments were conducted to test ideas for fossil fuel carbon dioxide ocean disposal as a sold hydrate at depths ranging from 349 to 3627 meters and from 8°to 1.6°C. Hydrate **formed** instantly from the gas phase at 349 meters but then **decomposed** rapidly in ambient seawater.*
> …
>
> *At 3 kilometers, you needed only 10 wells because the increased temperature **lower**ed the viscosity of the CO_2, allowed it to **slide** more easily into the reservoir.*

We can annotate a consequence event chain as shown in Figure 1 and a causal event chain in Figure 2. An example of the ontology markup is provided in the Appendix.

Event	Form
Object	***hydrate***
Origin	gas phase
Temperature	8-1.6 C
Depth	349 meters

→

Event	decompose
Object	***hydrate***
Environment	seawater

Figure 1: Example of Subsequence Event Chains

Event	lower
Object	***CO_2***
Agent	Increased temperature
Target	viscosity

→

Event	slide
Object	***CO_2***
Goal	reservoir
Volume	10 wells
Depth	3 kilometers

Figure 2. Example of Causal Event Chains

4. Carbon Sequestration Corpus

For the initial phase of this project we selected the carbon sequestration literature from isiknowledge.com. We have annotated 1397 journal article abstracts. On average each abstract includes 281 words.

In the future we plan to employ the large data repository proposed in (Lee et al., 2009). This repository will be constructed by domain scientists and engineers and includes various disciplines and media. Specifically, in the first year it will include 500MB theses and journal articles, 10MB progress reports and 50MB presentations and tutorials. After four years, the size of each genre is expected to reach 10 times of the current version.

5. Event Chain Annotation

Sustainable solutions of many annotation challenges for this carbon sequestration domain require highly interdisciplinary approaches. We have organized a team composing of both chemical engineers and computational linguists to conduct the corpus annotation.

We followed the general annotation scheme in the cross-document information extraction task that we proposed in (Ji et al., 2009). The detailed linguistic annotation steps are presented in the following subsections. Each step is done by two annotators independently and then adjudicated by a domain science expert for the final answer-key. In total it took one annotator about 20 minutes for each abstract. Our annotations are freely available for research purpose at http://nlp.cs.qc.cuny.edu/carbonie.html. The annotations will be gradually updated as the project progresses.

5.1 Predicate Annotation and Word Class Acquisition

During the first step, we followed the annotation guidelines in PropBank (Palmer et al., 2005; Xue and Palmer, 2009) by re-defining the collection of main verbs. We have defined 40 types of events in the domain of carbon sequestration, and followed an annotation guideline with structures similar to the ACE 2005 event extraction task (NIST, 2005). We also extensively exploited the manually constructed verb clusters such as VerbNet (Kipper et al., 2006). In addition, we applied the open-domain automatic verb clustering methods described in (Ji, 2009) to extend the coverage of verbs. For each Chinese verb in the semantic corpora such as PropBank, we search its aligned English words from the parallel corpora to construct a cluster including frequent English verbs. Then we can acquire Chinese verbs from the other direction and continue the iterations. For example, we can get the cluster of the "form" event with frequency information from small parallel corpora:

形成 → {found:198 set:183 form:43 create:4 launch:3 build:2 organize:2 forge:1 become:1}

Then we asked the annotators to filter the word alignment errors using verb lists and part-of-speech tagging.

5.2 Noun Phrase Chunking and Argument Labeling

It's also important to identify some domain-specific terminologies such as "computer modeling", "CO2 sequestration" and "power plants". We realized this by two steps of annotations. The first step is to extract all noun phrases from the sentences. We followed the guidelines of the Penn Treebank (Marcus et al., 2004) to annotate shallow parsing information for each article. For example, noun phrases are annotated for the following sentences in (Mancino and Reitenbach, 2008):

> *How can anyone make [NP data-driven decisions] about [NP the future] when we don't have access to [NP future data] ? [NP Computer modeling] is [NP our best-available option].*
>
> *Researchers are also studying [NP an area in Canada] where [NP hundreds of thousands of storage wells] have been exposed to [NP acid gas (CO 2 and hydrogen sulfide)]. Data from [NP so many wells] hold a wealth of information relevant to [NP CO 2 sequestration] and should generate statistics.*

In an offline procedure, a distributed version of K-means phrase clustering method (Lin and Wu, 2009) is applied to cluster the Google n-gram (n=5) corpus Version II, which can be viewed as a compressed summary of the web. Google n-gram Version II includes 207 billion tokens selected from the LDC-released Version I, consisted of 1.2 billion 5-grams extracted from about 9.7 billion sentences. All these 5-grams are automatically annotated with part-of-speech tags based on their original sentences. We then choose those phrases with high entropy of context as multi-word expressions. If a phrase is involved in any event, we try to assign a pre-defined role to it.

The inter-annotator agreement for this step is about 92%. The main disagreement is on whether a modifier word should be included in the argument phrase or not. For example, domain science annotators tend to include "deep" in "deep ocean" as an argument because they think "ocean" itself is not sufficient to represent domain-specific knowledge.

5.3 Temporal and Causal Relation Detection

Finally we link those relevant experiments if they are involved in temporal or causal relations. We wrote our annotation guideline for this step by merging the guidelines on the Time arguments in the ACE Event Corpus, Before/After annotations in TimeBank (Pustejovsky et al., 2003), ARGM-TMP relations in Propbank, Temporal Connective and Event-Event relation annotation in the UMD Semantic Annotation Corpus (Dorr and Onyshkevych, 2008). There were some

definition conflicts among these different guidelines, which have led to revisions and extensions of our annotations. The inter-annotator agreement for both relations was around 84%, which is surprisingly better than the news domain reported in (Ji et al., 2009). The main reason is that the experiments are usually described in a temporal order or indicated by some obvious subsequence words. More importantly, we found that a majority of the causal relations between experiments occur within single sentences, which has made the annotation much easier than the news domain. In the news domain, sometimes inferences across sentences or documents are needed in order to determine causal relations.

6. Related Work

There have been a lot of IE corpora developed for the bio-medical domain (e.g. Pyysalo et al., 2007), but to the best of our knowledge this is the first effort at annotating corpus for the domain of carbon sequestration.

Several recent studies have stressed the benefits of using unsupervised word or phrase clustering as additional knowledge to improve supervised learning. For example, Miller et al. (2004) proved that word clusters can significantly improve English name tagging. Ji (2009) used cross-lingual predicate cluster acquisition to improve bilingual event extraction in an inductive learning framework. Lin and Wu (2009) applied a web-scale phrase clustering algorithm to improve name tagging and query classification. Pantel and Lin (2003) described a clustering by committee algorithm to automatically discover word senses.

7. Conclusion and Future Work

We have described a project of annotating event chains for an important domain – carbon sequestration. We expect that our annotation resources will produce significant advances in inter-operability through new IE techniques and permit scientists to build knowledge that will provide better understanding of important scientific challenges in this domain, sharing and re-use of diverse data sets and experimental results in a more efficient manner.

Once the annotations are done for a significant amount of literature, we will aim to develop prototype models to automatically generate ontology, investigate domain adaptation techniques and conduct sublanguage analysis that includes automatic word class acquisition and pattern learning. We are also interested in applying the repeated active learning techniques as described in (Sheng et al., 2008) to speed up the annotation process. In addition, for this particular domain, some prestigious papers are published in foreign languages such as Japanese. Therefore in the future we also intend to extend this project to cross-lingual annotations in order to conduct effective information translation.

8. Appendix: Ontology Markup in Metadata

<event_chain ID="EV2" Centroid="CO_2 Geological Sequestration" Relation="Causal">
<extent START="263" END="427">At 3 kilometers, you needed only 10 wells because the increased temperature lowered the viscosity of the CO_2, allowed it to slide more easily into the reservoir. </extent>
<event id="EV2-1" TYPE="Lower">
<anchor START="341" END="345"> lower </anchor>
<argument id="E2-1" ROLE="Agent" START= "318" END="340">increased temperature </argument>
<argument id="E26-32" ROLE="Object" START= "370" END="372">CO_2</argument>
<argument id="E3-1" ROLE="Target" START= "353" END="361">viscosity</argument>
</event>
<event id="EV2-2" TYPE="Movement">
<anchor START="" END="">slide</anchor>
<argument id="E24-29" ROLE="Object"START= "370" END="372">CO_2</argument>
<argument id="E26-32" ROLE="Goal" START= "416" END="424">reservoir</argument>
<argument id="E36-2" ROLE="Volume" START="297" END="304">10 wells</argument>
<argument id="E41-3" ROLE="Depth" START= "267" END="278">3 kilometers</argument>
</event>
</event_chain>

9. Acknowledgements

This work was supported by the U.S. National Science Foundation Faculty Early Career Development (CAREER) Award under Grant IIS-0953149, the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, Google, Inc., CUNY Research Enhancement Program, Faculty Publication Program and GRTI Program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

10. References

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain Adaptation with Structural Correspondence Learning. *Proc. EMNLP 2006*.

Peter G. Brewer, Gernot Friederich, Edward t. Peltzer and Franklin M. Orr Jr. 1999. Direct Experiments on the Ocean Disposal of Fossil Fuel CO 2. *Science*.

Bonnie J. Dorr and Boyan A. Onyshkevych. 2008. From Linguistic Annotations to Knowledge Objects. *Proc. NSF Symposium on Semantic Knowledge Discovery: Organization and Use*, New York University, 2008.

Ralph Grishman. 2001. Adaptive Information Extraction and Sublanguage Analysis. *Proc. Workshop on Adaptive Text Extraction and Mining at Seventeenth International Joint Conference on Artificial Intelligence.*

Heng Ji. 2009. Unsupervised Cross-lingual Predicate Cluster Acquisition to Improve Bi-lingual Event Extraction. *Proc. HLT-NAACL 2009 Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics.*

Heng Ji, Ralph Grishman, Zheng Chen and Prashant Gupta. 2009. Cross-document Event Extraction and Tracking: Task, Evaluation, Techniques and Challenges. *Proc. Recent Advances in Natural Language Processing 2009.*

Jing Jiang and ChengXiang Zhai. "Instance weighting for domain adaptation in NLP." *Proc. ACL'07*.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extensive Classifications of English verbs. *Proc. the 12th EURALEX International Congress.*

Dekang Lin and Xiaoyun Wu. 2009. Phrase Clustering for Discriminative Learning. *Proc. ACL 2009.*

Dekang Lin, Ken Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, Kapil Dalwani and Sushant Narsale. 2010. New Data, Tags and Tools for Web-Scale N-grams. *Proc. LREC 2010*.

Jae W. Lee, Angelo Lucia, Jianting Zhang, Ann Zimmerman, DataNet: An Emerging Cyber-infrastructure for Sharing, Re-Using & Preserving Data for Scientific Discovery & Learning. 2009. *The American Institute of Chemical Engineers Journal (AIChe Journal), Wiley-Blackwell. (10.1002/aic.12085)*.

Anthony Mancino and Gail Reitenbach. 2008. The Future of Coal Power: Modeling Geological Sequestration of CO2. COALPOWER Magazine.

Mitchell P. Marcus, Mary Ann Marcinkiewicz and Beatrice Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. 2004. *Computational Linguistics*. 19(2). pp. 313-330.

Scott Miller, Jethran Guinness and Alex Zamanian.2004. Name Tagging with Word Clusters and Discriminative Training. *Proc. HLT-NAACL2004*. pp. 337-342. Boston, USA.

National Academy of Engineering. 2008. Grand Challenges for Engineering. http://www.engineeringchallenges.org/

NIST. 2005. Automatic Content Extraction 2005. http://www.itl.nist.gov/iad/mig/tests/ace/2008/.

Martha Palmer, Daniel Gildea and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*. Volume 31, Issue 1. pp. 71-106.

Patrick Pantel and Dekang Lin. Automatically Discovered Word Senses. *Proc. HLT-NAACL 2003*.

J. Pustejovsky, P.Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B.Sundheim, D. Day, L. Ferro and M. Lazo. 2003. The Timebank Corpus. *Corpus Linguistics*. pp. 647-656.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Bjorne, Jorma Boberg, Jouni Jarvinen and Tapio Salakoski. 2007. BioInfer: a Corpus for Information Extraction in the Biomedical Domain. *BMC Bioinformatics*.

V. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. *Proc. the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08), pages 614–622, 2008.*

Nianwen Xue and Martha Palmer. 2009. Adding semantic roles to the Chinese Treebank. *Natural Language Engineering*, 15(1):143-172.