LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-TR-209869

A Novel Approach to Semantic and Coreference Annotation at LLNL

M. Firpo

February 18, 2005

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

A Novel Approach to Semantic and Coreference Annotation at LLNL

Michael Firpo

Abstract

A case is made for the importance of high quality semantic and coreference annotation. The challenges of providing such annotation are described. Asperger's Syndrome is introduced, and the connections are drawn between the needs of text annotation and the abilities of persons with Asperger's Syndrome to meet those needs. Finally, a pilot program is recommended wherein semantic annotation is performed by people with Asperger's Syndrome.

Executive Summary

The primary points embodied in this paper are as follows:

- Document annotation is essential to the Natural Language Processing (NLP) projects at Lawrence Livermore National Laboratory (LLNL);
- LLNL does not currently have a system in place to meet its need for text annotation;
- Text annotation is challenging for a variety of reasons, many related to its very rote nature;
- Persons with Asperger's Syndrome are particularly skilled at rote verbal tasks, and behavioral experts agree that they would excel at text annotation;
- A pilot study is recommend in which two to three people with Asperger's Syndrome annotate documents and then the quality and throughput of their work is evaluated relative to that of their neuro-typical peers.

1. Introduction

1.1 Purpose

The purpose of this document is to a) describe the necessity of high quality text annotation to the field of Natural Language Processing (NLP), b) indicate the challenges in providing such annotation, c) introduce Asperger's Syndrome (AS) and indicate the advantages persons with AS would have in performing text annotation, and d) recommend a pilot study to have persons with AS annotate documents and compare their quality and throughput with that of neuro-typical annotators.

1.2 Scope

For the scope of this document, NLP will refer to the processing of free-text. In other words, it will consider neither spoken language nor structured or semi-structured text. Furthermore, the proposed technology is intended for use with English language texts only. The word *corpus* (and its plural *corpora*) refers to a body of text documents, typically in plain text or HTML formats. Finally, the words annotation and tag are used interchangeably throughout this document.

1.3 Motivation for having Annotations

Two of the primary tasks in NLP are *Information Extraction* (IE) and *Discourse Integration* (DI). In IE, the computer identifies important information within a body of text; in DI, multiple references and other confusing (to the machine) aspects of the texts are resolved.

These tasks are usually managed by using either an expert system or an adaptive (or machine learning) system. Expert systems rely on an expert user to provide needed intelligence for solving the problem; the expert essentially provides the rules used by the system to implement IE or DI. Adaptive systems do not require an expert [1, 2]; the needed intelligence is provided by training data – pretagged corpora from which the adaptive system learns a set of rules for implementing IE or DI.

Adaptive systems require thousands of documents for training. Expert systems require on the order of hundreds of documents for testing.

Furthermore, DI and various aspects of IE (Entity Extraction, Attribute Extraction, Relationship Extraction, and Event Extraction) require different sorts of annotations. So the same corpus may, in fact, need to be tagged multiple times for multiple applications.

According to most commercial vendors, expert systems have out-performed adaptive systems. However, adaptive systems have the advantage that they do not require an expert. Academics, including Andrew McCallum [3] of the University of Massachusetts, generally believe that adaptive systems are better suited to solve these sorts of problems in the long term. I contend that, while expert systems are currently more commercially viable, the adaptive systems will provide a better solution to the needs of LLNL scientists and engineers, over time.

In addition to adaptive and expert systems, hybrid systems are sometimes used to accomplish IE and DI. The expert has a reduced role in hybrid systems compared with expert systems. In a hybrid system, the training data is used for the machine learning component of rule generation, and then the expert simply adjusts the rules, rather than generating them from scratch. However these systems have their limitations as well. Humans tend to think about rules differently than computers do. Humans prefer high coverage – that is, rules that cover a large number of cases. A typical expert will look for a rule that covers 70–90% of cases, and then another that covers 70–90% of the remaining cases. By contrast, a computer will typically find a

large number of rules, each of which provides 5–10% coverage. For this reason, it can be difficult for the human to work with computer-generated rules.

Finally, there are various systems and methods in development that are designed to reduce the size of the annotation task: bootstrapping methods that make use of redundant information; active learning systems that 'know what they don't know' and request input from the user on the cases they don't know; and there is even a system designed to eliminate the human tagger by generating extraction patterns (rules) from untagged text, seeded simply with a set of keywords that the user expects to find within the corpus [4]. Perhaps the most promising system is one described by Rosie Jones, et. al. [5] that combines the bootstrapping with the active learning. The active learning aspect keeps the bootstrap on track, while the bootstrap reduces the annotation workload considerably. These systems vary in their accuracy and usefulness. In general the less input received from the user, the greater the chances are that an automatic code can start bootstrapping on a tangent and end with a lot of useless annotations.

Despite the progress being made, the need for a certain amount of annotation will remain until a superior method is fully developed.

1.4 Advantage of Alternate Perspectives

During the Vietnam War, the U.S. Military used colorblind people to find camouflaged forces. When the enemy hid under large tarps that matched the color of the surrounding foliage, typically color-sighted people were thrown off. The colorblind people were far more perceptive to the differences in texture between the tarps and the foliage. Having a unique perspective made these people better suited to the task at hand.

In the same way, having a unique perspective on the information contained in free-text may make people with Asperger's Syndrome better suited to text annotation than their neuro-typical peers.

2. Description of Annotation

2.1 Elementary Semantic Annotation

The goal of Information Extraction is for the computer program to identify, within a corpus, words and phrases that carry important meaning – important, that is, to the user. The user defines what sorts of meaning are important in a separate document called an *ontology* – a specification of semantic constructs of interest. Currently, most ontologies are written in some form of XML. Three commonly used formats that have been developed for use in ontologies are RDF/RDFS (Resource Description Framework/Schema), DAML (DARPA Agent Markup Language), and Web Ontology Language (OWL). An example ontology in DAML format is provided in the appendix.

Semantic (or Sense) Annotation is used for testing and often for training of IE tools in NLP. The human annotator, or tagger, identifies the meaningful words or phrases one would expect the IE tool to identify. Typically, this is done by employing an annotation tool that does one of two things: either it incorporates XML style tags directly into the original text, or it saves a meta-file with references back to the positions of words and phrases in the original document. The following is an example of what semantically tagged text looks like with color-coded XML tags:

> <victim>15 people</victim> have been <care>admitted to hospital</care> in <location>Kuzbass [Kemerovo Region]</location> with suspected <condition>tickborne encephalitis</condition>, the <actor>Kemerovo Regional Administration's press service</actor> told RIA on <date>Mon 17 May 2004</date>, referring to the latest figures supplied by the <actor>State Health Inspectorate's Regional Centre</actor>. Over <victim>3000 local people</victim> have already sought <care>medical help</care> for tick bites. Out of that 3000, over 2000 have done so during the past week [<date>mid May 2004</date>].[1]

2.2 Elementary Coreference Annotation

Discourse Integration involves connecting meaning between sentences, handling coreference, and managing other idiosyncrasies of the English language. These characteristics may make the English language more interesting for humans, but they create difficulties for machines.

The most challenging aspect of DI is the *coreference* problem.

The coreference problem can be illustrated with the following sample discourse:

> George W. Bush invited Diane Feinstein to tea.
> She accepted his invitation.
> The President was happy to see the Senator.

In this text sample, two people are referenced by name, by pronoun, and finally by description. The fluent human reader is automatically aware that "Diane Feinstein", "she", and "the Senator" all refer to the same individual. The challenge of coreference is to teach the computer to recognize these truths, even in more complicated cases.

Coreference tagging is used for testing and sometimes for training of DI tools. A human tagger may tag the above example for coreference in the following way:

> <person ID=1>George W. Bush</person> invited <person ID=2>Diane Feinstein</person> to tea.
> <person ID=2>She</person> accepted <person ID=1>his</person> invitation.

[1] The text comes from ProMed; the ontology is my own.

<person ID=1>The president</person> was happy to see <person ID=2>the senator</person>.

3. Current Work at LLNL

The Information Operations and Analysis (IOA) Center's Louisiana Project currently performs semantic and coreference tagging. Professional Computer Scientists (285.0s) break from their high-priority software development tasks to sit and stare at the computer screen and read articles in a corpus. They select important words and phrases and then specify the significance, as defined in the working ontology.

The work is very simple and doesn't require a software engineer or even a person with a high school diploma.

One concern related to semantic and coreference tagging is that, as the task loses its novelty, the annotator loses interest, and then the quality of annotation suffers. Particularly if the annotator is over-qualified, he or she may think that they could be doing more intellectually stimulating work, and the quality can take another hit.

4. Challenges of High Quality Semantic and Coreference Tagging

Document annotation is challenging for the following reasons:

- Large volumes of data are required.
- The process is tedious.
- Though the process is error prone, high accuracy is important to subsequent steps of NLP.
- Even the most minute details must be handled consistently:
  - whether to tag the article preceding a noun;
  - how to handle phrasal verbs (e.g. "he made up the assignment" vs. "he made the assignment up");
  - which prepositional phrases to include in an event;
  - etc.
- The same text may need to be tagged more than once because tagging requirements differ for different tasks (Entity Extraction, Event Extraction, Coreference Resolution).

5. Asperger's Syndrome

As mentioned in section 4, there are challenges associated with tagging: it can be boring and tedious, yet it requires intensive attention to detail. Unfortunately, with tedium generally comes a decreased attention to detail.

There is a group of people who are well suited to this type of work. There is a high-functioning form of autism called Asperger's Syndrome (AS). People with AS are characterized as having above-average intelligence and excellent rote verbal skills. AS people can get involved in a rote task (like tagging corpora) and give it their undivided attention for hours. This group of people has the patience for very rote work, and they do it very well.

According to the Diagnostic and Statistical Manual of the American Psychiatric Association (1994), "Asperger's Disorder patients have a preoccupation with parts of objects." This is a skill that tagging requires – it is seeing not only the trees in the forest, but concentrating on the individual leaves more than the larger whole.

Dr. Sandie Frawley, Director of the Center for Educational and Psychological Services in San Ramon; and William Shryer, LCSW, BCD, Clinical Director of Diablo Behavioral Health Care in Danville, are experts in the area of AS. Both contend that AS people would be well suited to the task of tagging corpora. Shryer works with adults who have AS. After hearing a description of the task of tagging a corpus as, "sitting in front of a computer reading articles all day, highlighting the bits that are interesting," his immediate response was, "That would be perfect for [someone with AS]; that is how their brains are wired."

Of course, if there were no downside to AS, it wouldn't be considered a syndrome. AS people tend to lack pragmatic skills. Fortunately, in this particular application, pragmatics is not a concern. If potential coworkers know what to expect and have communication tools, then it would be similar to working with anybody else. Some people with AS also have attention deficits or oppositional difficulties. However, these qualities occur in the minority of the AS population. Shryer says that there are many AS people in the area who would be able to successfully perform such work, and who would be free from attention and oppositional problems.

People with AS would be able to provide high-quality annotation because they thrive in an environment that involves rote tasks. They prefer to be asked questions with obvious answers because they can be correct, and then they feel good about themselves. They are comfortable devoting long periods of time to repetitious tasks. Annotating text would provide an excellent opportunity for AS people to sit and get many things right in rapid succession.

6. Recommendation

I recommend a pilot program wherein persons with Asperger's Syndrome are contracted to perform semantic tagging of corpora.

We (representatives of LLNL conducting the pilot) will identify a) an unclassified corpus, b) an ontology, and c) the tagging specifics of a target application, which will be the recipient of the resulting tagged corpus. Then we will work with William

Shryer to identify two or three persons with AS who would be well suited to the task. Then arrangements will be made for them to tag corpora for 40 hours per week.

Funding for such a pilot might come through the National Institute of Mental Health (NIHM) or a related State agency.

If the pilot is successful, a team of taggers would be formed. Then we would be prepared when a corpus of 1,000,000 words needs to be tagged in three different ways for three different applications. The work would be given to an experienced and appropriately trained team. The taggers would be compensated in a manner more generous than other alternatives available to them, at a substantial savings to LLNL over having the work performed by software developers. The annotations would be performed more accurately and with increased throughput than is presently available.

7. Appendix

Below is an example ontology in DAML format, which is built on top of RDF:

```
<?xml version='1.0' encoding='ISO-8859-1'?>

<!DOCTYPE rdf:RDF [
   <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#">
   <!ENTITY rdfs "http://www.w3.org/2000/01/PR-rdf-schema-
19990303#">
   <!ENTITY daml 'http://www.daml.org/2001/03/daml+oil#'>
]>


<rdf:RDF
  xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns = "http://www.daml.org/2001/03/daml+oil#"
  xmlns:daml = "http://www.daml.org/2001/03/daml+oil#"
  xmlns:base =
"http://www.cs.umd.edu/projects/plus/DAML/onts/base1.0.daml#"
>

<Ontology rdf:about="">
   <versionInfo>Firpo-nonsense-ontology, v.0.1</versionInfo>
   <comment>My first shot at a DAML ontology.</comment>
</Ontology>

<!-- Classes -->

<Class rdf:ID="Party">
  <comment>Base party class</comment>
</Class>

<Class rdf:ID="Shower">
  <subClassOf rdf:resource="#Party"/>
```

```
</Class>

<Class rdf:ID="Reception">
  <subClassOf rdf:resource="#Party"/>
</Class>

<Class rdf:ID="Holiday">
  <subClassOf rdf:resource="#Party"/>
</Class>

<Class rdf:ID="HouseWarming">
  <subClassOf rdf:resource="#Party"/>
</Class>

<Class rdf:ID="Venue">
  <comment>address</comment>
</Class>

<Class rdf:ID="Guest">
  <comment>person attending the party</comment>
</Class>

<Class rdf:ID="Time">
  <comment>a point in time</comment>
</Class>

<Class rdf:ID="Honoree">
  <comment>honoree</comment>
  <subClassOf rdf:resource="#Guest"/>
</Class>

<Class rdf:ID="Date">
  <comment>date</comment>
</Class>


<!-- Relationships -->

<Property rdf:ID="on-date">
  <label>is on</label>
  <domain rdf:resource="#Party"/>
  <range rdf:resource="#Date"/>
</Property>

<Property rdf:ID="begin-time">
  <label>starts at</label>
  <domain rdf:resource="#Party"/>
  <domain rdf:resource="#Reception"/>
  <domain rdf:resource="#Shower"/>
  <domain rdf:resource="#Housewarming"/>
  <domain rdf:resource="#Holiday"/>
  <range rdf:resource="#Time"/>
```

```
</Property>

<Property rdf:ID="end-time">
  <label>ends at</label>
  <domain rdf:resource="#Party"/>
  <domain rdf:resource="#Reception"/>
  <domain rdf:resource="#Shower"/>
  <domain rdf:resource="#Housewarming"/>
  <domain rdf:resource="#Holiday"/>
  <range rdf:resource="#Time"/>
</Property>

<Property rdf:ID="at-venue">
  <label>is at</label>
  <domain rdf:resource="#Party"/>
  <domain rdf:resource="#Reception"/>
  <domain rdf:resource="#Shower"/>
  <domain rdf:resource="#Housewarming"/>
  <domain rdf:resource="#Holiday"/>
  <range rdf:resource="#Venue"/>
</Property>

<Property rdf:ID="honoree">
  <label>in honor of</label>
  <domain rdf:resource="#Party"/>
  <domain rdf:resource="#Reception"/>
  <domain rdf:resource="#Shower"/>
  <range rdf:resource="#Honoree"/>
</Property>

<Property rdf:ID="host">
  <label>throws</label>
  <domain rdf:resource="#Party"/>
  <domain rdf:resource="#Reception"/>
  <domain rdf:resource="#Shower"/>
  <domain rdf:resource="#Housewarming"/>
  <domain rdf:resource="#Holiday"/>
  <range rdf:resource="#Guest"/>
</Property>

<Property rdf:ID="attender">
  <label>attends</label>
  <domain rdf:resource="#Party"/>
  <domain rdf:resource="#Reception"/>
  <domain rdf:resource="#Shower"/>
  <domain rdf:resource="#Housewarming"/>
  <domain rdf:resource="#Holiday"/>
  <range rdf:resource="#Guest"/>
</Property>

</rdf:RDF>
```

8. References

1. Ciravegna, Fabio, "Adaptive Information Extraction from Text by Rule Induction and Generalisation," *Proc. 17th International Joint Conference on Artificial Intelligence (IJCAI 2001),* Seattle, WA, 2001.

2. Kushmerick, Nicholas, and Bernd Thomas, "Adaptive Information Extraction: Core Technologies for Information Agents," *Intelligent Information Agents R&D in Europe: An AgentLink Perspective*, 2002.

3. McCallum, Andrew. Conversation held between Andrew McCallum, Anton Fulmen, Sarah Tyler, and Michael Firpo, November 16, 2004.

4. Riloff, Ellen, "Automatically Generating Extraction Patterns from Untagged Text," *Proc. 13th National Conference on Artificial Intelligence (AAAI-96),* pp. 1044–1049, AAAI Press, 1996.

5. Jones, R., R. Ghani, T. Mitchell, and E. Riloff, "Active Learning for Information Extraction with Multiple View Feature Sets," *ECML-03 Workshop on Adaptive Text Extraction and Mining*, 2003.