Evaluation Plan

Anthony Gentile
Lisa Gress

Formalism of GREC semantic representation

The semantic annotations of the GREC corpus are centered around events related to gene expression and regulation. These events stem from verbs or nominalized verbs within the MEDLINE abstract and title. The first sentence is evaluated to see if it is on the topic gene regulation, if not, it goes on to the next sentence to do the same check. Of these sentences, the main verb of the sentence is located. This is the verb that describes the main event of the sentence. This event is annotated first, and then events from other verbs in the sentence thereafter. This means that events are not necessarily annotated in order. All events for all verbs in a sentence are annotated as long as the sentence is related to gene regulation.

Related variables for these events are identified and then semantic roles are listed corresponding to these variables. It is possible for the identified semantic roles to correspond with another event as opposed to an identified trigger word. This means that there can be nested events. These annotations are available in an XML format as well as a standoff format.

Different types of information encoded in the formalism

Within the GREC annotations, we are provided with event relationships from verbs. Within these we have two main types of information. First, we have concept classes used to label biological processes such as Repressor, Regulator, and Gene. Second, we have semantic role types such as Theme, Manner, Agent, and Location. These two types of information are tied to spans of the sentence by unique identifiers and structured within particular verb events that help us identify relationships.

Atomic pieces of the representations

Using the standoff format of the annotations, the atomic pieces can be easily identified.

For each MEDLINE abstract we have three files corresponding to the PMID of the abstract.

```
1885551.txt
1885551.a1
1885551.a2
```

The .txt file contains the text of the MEDLINE abstract on two lines. The first being the abstract title, the second being the body of the abstract.

The .a1 file contains named entity and event argument text spans. The .a2 file contains annotations relating to events.

1885551.a1 (snippet)

```
T1 Repressor 0 23 Integration host factor
T2 SPAN 30 42 specifically
T3 Locus 46 60 multiple sites
T4 Transcription 115 128 transcription
T5 Regulator 181 202 a DNA-binding protein
T6 Gene 224 228 gene
```

For each line we start with an ID for the annotation, followed by the annotation type, character offsets, and finally the surface text span of the annotation.

1885551.a2 (snippet)

```
T26 GRE 24 29 binds
T27 GRE 106 114 inhibits
T28 GRE 208 220 participates
T29 Regulation 229 239 regulation
T30 GRE 527 532 binds
T31 Transcription 586 599 transcription
T35 GRE 825 832 binding
E1 GRE:T26 Agent:T1 Manner:T2 Destination:T3
E2 GRE:T27 Agent:T1 Theme:T4
E3 GRE:T28 Agent:T5 Location:T9,T11,T12 Descriptive-Agent:E4,T7,T8
E4 Regulation:T29 Theme:T6 Location:T9
```

At the top of the .a2 files are annotations related to event trigger words. These are in the same format as .a1 files and often of the type GRE (Gene Regulation Event) which are often assigned to the top level verb trigger words. After these event trigger word annotations, we have event annotations, for which the IDs start with 'E' instead of 'T'. Following this identifier is an event type and argument identifier pair followed by a list of semantic role and argument identifier pairs for the corresponding event.

The atomic pieces to be used

We will use most of this information in our mapping process, however, we do not plan to consider the type information, such as GRE, Regulation, and other biological concept types in our work. We will be focused solely on the semantic role information provided. Additionally, we will need to do certain manipulations. One such manipulation will be to change offsets from being on the abstract level to a sentence level to mimic the MRS.

Calculation of precision and recall

We hope to evaluate the MEDLINE abstract sentences independently using MRS and also with our gold standard GREC annotations. By doing so we will be able to identify the sentences in which

MRS and GREC annotations representations deviate and further delve into what those deviations are. We will measure the precision by taking the ouput of our mapping of the MRS format to GREC format and seeing where we have information that is not in the GREC annotations (gold standard). For recall, we will take the the output of our mapping to see which parts of the GREC annotations (gold standard) are not mapped to correctly.

An example of such calculations are as follows:

```
GREC
"transcription" <9:22>
ARG1: Manner <0:8>
ARG2: Source <28:51>

MRS
"_transcription_n_1_rel" <9:22>
ARG0: x3 <9:22>
```

From the MRS, the `"_transcription_n_1_rel"` relation doesn't take any arguments so it doesn't provide us any relationships, whereas the GREC annotation (gold standard) provides for two. Given this sole context, we would have a precision of 0/2 and a recall of 0/2.

Calculation of elements not mapped

In addition to quantifying instances, we also intend to list the deviations for inspection. We plan to count and list which parts of the MRS were not needed in the mapping procedure for futher analysis. We hope to be able to provide a similar type of precision and recall calculation for this information as well, but are hestitant to show an example as we have yet to determine how we intend to structure that output.

References

Paul Thompson, Syed A Iqbal, John McNaught and Sophia Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10(1):349.