Target Corpus Description

Anthony Gentile
Lisa Gress

We will be examining the Gene Regulation Event Corpus (GREC) which is comprised of MEDLINE abstracts. MEDLINE is a database maintained by the National Library of Medicine and it contains citations and abstracts from biomedical journals. Of the 240 MEDLINE abstracts in the GREC, 167 of these abstracts are on the subject of the *E. coli* species, while the remaining 73 are on the subject of the human species. The purpose of this corpus is for training information extraction tools, particularly those attempting to extract events from the biomedical domain. The corpus has been used to produce semantic frames for BioLexicon, a terminological resource used for biomedical text mining. The corpus and associated annotation guidelines are available for download at `http://www.nactem.ac.uk/GREC/`. The annotation scheme and annotated corpus can be cited using Thompson et al. (2009). The individual source texts are abstracts of journal articles and can be cited individually, if need be.

Annotation scheme

The abstracts are annotated with regard to gene regulation and expression events, defined by Thompson et al. (2009) as "events that describe any interaction which leads, either directly or indirectly, to the production of a protein." Only sentences that contained a description of a transcription, translation, or post-transcriptional modifications were selected for annotation.

Events related to gene expression and regulation are comprised of either verbs (such as *transcribe* or *regulate*) or nominalized verbs (such as *transcription* or *regulation*). A total of 3067 events were annotated in this corpus. For each event, the authors seek to identify all structurally-related arguments in the same sentence. Arguments can include information such as location, manner, timing, and condition (Thompson et al., 2009). Each argument is assigned one of 13 semantic roles. In addition to a semantic role, an argument may also be assigned a biological concept type that is specific to the gene regulation domain.

In addition to semantic roles that were selected (and sometimes modified to fit the domain) from VerbNet and PropBank, a few domain-specific roles were created. The 13 semantic roles used in annotation are AGENT, THEME, MANNER, INSTRUMENT, LOCATION, SOURCE, DESTINATION, RATE, TEMPORAL, CONDITION, PURPOSE, DESCRIPTIVE-AGENT, and DESCRIPTIVE-THEME.

Biological concept labels were chosen from five different hierarchies that were based on the Gene Regulation Ontology. The five hierarchies were *Nucleic_Acids*, *Proteins*, *Living_Systems*, *Processes*, and *Experimental*. Each hierarchy consisted of increasingly specific labels. For each argument, the most specific possible label was chosen from the hierarchies, based on context.

Thompson et al. (2009) give the following example of a labeled event (shown in example 1), in which the event is the verb *activated* and the arguments of the event are *In Escherichia Coli*,

glnAP2, and *NifA*. These arguments were labeled with semantic roles and biological concept types as shown in Table 1.

(1) In Escherichia Coli, glnAP2 may be **activated** by NifA.

Argument	**Semantic role**	**Biological concept**
NifA	AGENT	Activator
glnAP2	THEME	Gene
In Escherichia Coli	LOCATION	Wild_Type_Bacteria

Table 1: Arguments of example sentence labeled with semantic roles and biological concepts

Linguistic constraints

In order to ensure that the argument text spans chosen for annotation were consistent, the text was tagged using the GENIA tagger prior to annotation. Example 2 shows a GENIA-tagged sentence (Thompson et al., 2009) Only base NP chunks were annotated and any NP with additional descriptive information, typically indicated by an NP following a preposition, was excluded. Thus, the only arguments of the event *encoded* in example 2 are *The klebsiella rcsA gene* and *a polypeptide*.

(2) [NP The klebsiella rcsA gene] [VP encoded] [NP a polypeptide] [PP of] [NP 23 kDa].

Inter-annotator agreement

The annotation of the corpus was done by hand using a customizable annotation tool called WordFreak. Annotation was carried out by 6 biology PhD students with experience in gene regulation. Inter-annotator agreement (IAA) was calculated using F-Score for eight different subtasks: event identification, argument identification (split into relaxed span matches and exact span matches), semantic role assignment, biological concept identification, biological concept category assignment (split into exact category matches, parent category matches, and supercategory matches). When calculating IAA between two annotators, one annotation set was treated as the gold standard for computing precision and recall for the F-Score.

There was an initial period of annotation training, split into five cycles, during which IAA scores were calculated. The annotators were given feedback regarding their annotations after each cycle. The first four cycles were used to annotate abstracts concerning *E. coli* and the fifth cycle was used to annotate abstracts concerning humans.

After the training period, the final corpus was produced. Average IAA F-Score per subtask are given, split by species type, for the final corpus. For the *E. coli* species, scores range from a low of 71.02% for exact biological concept category assignment to a high of 95.52% for assignment of supercategory of biological concept. For the human species related abstracts, scores range from 66.03% for exact biological concept category assignment to 94.75% for assignment of supercategory of biological concept. IAA scores for semantic role assignment was around 88% for both the *E. coli* and human categories (Thompson et al, 2009).

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.