

BELA Corpus

Corpus overview

The BELA (Bio-Event Linguistic Annotation) corpus is a corpus of MEDLINE abstracts on the subject of E.Coli. The corpus has been annotated with bio-events relating to *gene regulation*, with a specific view to the acquisition of semantic frames. Within the BOOTStrep project, the corpus was used to acquire the semantic frames for verbs within the BioLexicon.

The annotation, which was undertaken by 7 PhD students at the University of Manchester, consisted of identifying gene regulation events described both by verbs and nominalised verbs, such as *transcription* or *expression*. For each relevant event, the following steps took place:

- All semantic arguments within the same sentence were identified
- Each semantic argument was assigned a semantic role from a set of 13 event-independent roles tailored to the domain.
- Where appropriate, arguments (or parts of arguments) were also assigned named entity/biological concept types, from a set of 61 categories which are tuned to the gene regulation domain and are mappable to the Gene Regulation Ontology (Splendiani et al, 2007), The types are organized into four entity-specific super-classes (DNA, PROTEIN, EXPERIMENTAL and ORGANISMS), and one event-specific super-class (PROCESSES).

It should be noted that the annotation does not correspond exactly to event instance annotation, due to the specific goal of acquiring semantic frames. Thus, for example, there were special rules for annotating lists, so that if a semantic argument consists of a list of entities of the same NE type, then the annotated span should only consist of *one* entity in the list. However, if the list contains items of differing NE types, then one item of *each* type should be annotated.

Further details of the annotation scheme may be found in the annotation guidelines.

The corpus is divided into 2 parts:

- 1) 597 abstracts, each annotated by a single annotator, containing a total of 3612 events
2. 80 pairs of double-annotated documents allowing inter-annotator agreement and consistency, and containing 1158 distinct events.

Analysis of the double-annotated abstracts revealed that inter-annotator agreement rates for various annotation subtasks lie in the range of 49- 78%. Further details may be found in Thompson et al. (2008)

Annotation tool and corpus format

Annotation was carried out using a version of the WordFreak tool (Morton & LaCivita, 2003) which was customised by ILC-CNR. The corpus consists of 2 types of files:

- 1) Text files containing the abstract text
- 2) Standoff XML annotation files produced by WordFreak.

Each file is named according to the PMID of the abstract.

The corpus consists of 2 directories:

- 1) The “BELA” directory contains abstracts annotated by a single annotator. In this directory, the text files are called <PMID>.txt and the annotation files are called <PMID>.txt.ann.
- 2) The “BELA-double-annotated” directory contains abstracts annotated by 2 or more annotators. In this directory, the text files are called <PMID>-<Annotator_Initials>-PD.txt and the annotation files are called <PMID>-<Annotator_Initials>-PD.txt.ann

The customised version of the tool is provided as a jar file with the corpus (wf_1_3.jar) together with a project file (BELA.prj). Opening the project file within the tool will allow the annotated files to be viewed.

The XML annotation files contain linguistic information obtained through pre-processing with the GENIA tagger (i.e. part-of-speech categories and syntactic chunk types), in addition to the manually added annotations. There are 3 main types of elements within these files, i.e. <Annotation>, <Relation> and <Argument>

The <Annotation> elements correspond to textual spans. The spans correspond both to individual tokens and to syntactic chunks. Each <Annotation> element has the following attributes:

- `annotator` – value is “tagger” if added through automatic annotation or the name of the annotator if done through manual annotation
- `confidence` – not used but defaults to “1.0”
- `id` – A unique id for the annotation element
- `category` – The named entity/biological concept type assigned to this span (if one has been assigned). Assignment of NE categories is only possible to NP, VP and VP-BIO chunks. VP-BIO chunks correspond to chunks which denote or potentially denote events.
- `span` – the start and end offsets of the text span corresponding to this annotation, separated by “. . .”
- `type` – Corresponds either to Part-of-Speech category of the syntactic chunk type of the annotation, depending on the type of span corresponding to the annotation

The <Relation> elements correspond to the annotated events. Each <Relation> element has 2 attributes, as follows:

- `anchor` – an id corresponding to the VP-BIO chunk on which the event is centred
- `type` – This is always “BEA” (for Bio-event annotation).

Each <Relation> element contains <Argument> elements, corresponding to the arguments of the event. These have 2 attributes, i.e.

- `id` – the id of the chunk corresponding to the argument (if one has been annotated)
- `role` - the semantic role assigned to the argument.

Note that there is always an <Argument> element which has the `role` value of “Verb”, in addition to the actual semantic arguments of the event. The id of this element corresponds to the actual verb (or nominalised verb) on which the event is centred, as opposed to the complete chunk over which the event was created.

References

Morton T. & LaCivita J. (2003). Word-Freak: an open tool for linguistic annotation. In *Proceedings of HLT/NAACL-2003*, pp 17--18.

Splendiani, A., Beisswanger, E., Kim, J-J., Lee, V. , Dameron, O. & Rebholz-Schuhmann, D. (2007) Bio-Ontologies in the context of the BOOTStrep project. Bio-Ontologies SIG Workshop, Vienna.

Thompson, P., Cotter, P., Ananiadou, S., McNaught, J., Montemagni, S., Trabucco, A. and Venturi, G. (2008). "Building a Bio-Event Annotated Corpus for the Acquisition of Semantic Frames from Biomedical Corpora". *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.