GENIA EVENT CORPUS

1 BASIC INFORMATION

1.1 Corpus composition

The corpus consists of 1000 MEDLINE abstracts. It is a subset of the original GENIA corpus, which was selected using the three MeSH terms *human*, *blood cells* and *transcription factors*.

1.2 Representation of the corpus (flat files, database, markup)

The corpus is provided as a set of XML files, which contain both the abstract text and annotations.

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Sophia Ananiadou
Address: Manchester Interdisciplinary Biocentre,131 Princess Street, Manchester M1 7DN, IK
Affiliation: National Centre for Text Mining, School of Computer Science, University of Manchester
Position: Director
Telephone: +44 161 306 3092
Fax: +44 161 306 5201
e-mail: Sophia.Ananiadou@manchester.ac.uk

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource is available on the MetaShare platform as an archive.

2.3 Copyright statement and information on IPR

The resource is freely available for research purposes

3 TECHNICAL INFORMATION

3.1 Directories and files

The archive contains the directory *Genia_Metaknowledge_Corpus.* It contains the following sub-directories and files:

- *Corpus* – directory containing 1,000 XML files. Each file contains one Medline abstract.
- *ModifiedGENIAtypes* – directory containing the DTD and CSS files for the XML files constituting the GENIA event corpus. All the XML files in the corpus are validated against the DTD. The CSS well works with the Opera web browser.
- *GENIAontologies* – directory containing two GENIA ontologies encoded in OWL. The GENIAterm40.owl defines the term classes on which the GENIA term annotation is based. The GENIAevent.owl defines the event classes on which the GENIA event annotation is based.
- *Guidelines_for_event_annotation.pdf* – Annotation guidelines used to produce the event annotations (see below for more details)
- *Meta-knowledge_annotation_guidelines.pdf* – Annotation guidelines used to add meta-knowledge annotation to the events (see below for more details)
- *README.html* – Provides basic introductory information about the corpus.

3.2 Data structure of an entry

This is not relevant as the corpus is a set of text files.

3.3 Resource size (nmb. of tokens, MB occupied on disk)

The corpus contains approximately 220,000 tokens. It requires approximately 21.7 MB of disk space.

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is a monolingual, annotated corpus.

4.2 The natural language(s) of the corpus

The language of the corpus is English.

4. 3 Domain(s)/register(s) of the corpus

The corpus contains abstracts of biomedical research articles.

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

The corpus (Kim et al, 2008) contains structural annotations obtained from MEDLINE, e.g. to identify the abstract title and main body of the abstract. The text is split into sentences. In each sentence, three types of information are annotated:

- Firstly, biomedical terms are identified and assigned categories from the GENIA term ontology.
- Secondly, event structures are identified and assigned categories from the GENIA event ontology.
- Thirdly, detailed information is annotated about how the event should be interpreted, according to its textual context. We call this information *meta-knowledge.*

Consider the following sentence:

The results suggest that LMP1 activates NF-kappa B

Term annotation identifies *LMP1* and *NF-kappa B* as terms, while event annotation identifies that there is a relationship between these terms, i.e. they participate in an event of type *Positive_Regulation*, in which *activates* is the *event trigger, LMP1* is the CAUSE of the event and *NF-kappa B* is the THEME, i.e. what is affected by the event. Finally, meta-knowledge annotation encodes that fact the event is a slightly speculated analysis of results rather than, e.g., a definite fact.

Meta-knowledge is classified according to 5 different dimensions:

- *Knowledge Type* – general information content of the event. Does it represent an investigation, observation, analysis, etc.
- *Certainly level* – the level of certainty associated with the occurrence of the event
- *Polarity* – whether or not the event is negated
- *Manner* - the rate, level, strength or intensity of the event (in biological terms)
- *Source* - the source or origin of the knowledge being expressed by the event. Specifically, we distinguish between events that can be attributed to the current study, and those that are attributed to other studies

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

Term annotations were manually annotated as part of the original GENIA annotation (see Kim et al., 2003). Terms were assigned

categories from the GENIA ontology (35 biologically categories corresponding to terminal nodes in the ontology). The term annotation takes care of semi-structured coordinated clauses by recovering ellipsis. As an example, the phrase *CD2 and CD 25 receptors* refers to two terms, i.e. *CD2 receptors* and *CD25 receptors*, but *CD2 receptors* doesn't appear in the text. The annotation is carried out in such a way to allow these separate terms to be identified. Term annotations are inline, within eaxh sentence annotation. Term annotation is described in more detail in Kim et al. (2003). See also http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=Technical+Term+Annotation.

Event annotation was carried manually out on top of term annotation to identify the biological processes in which the terms participate. Event annotations are attached to each sentence, providing information about the structure of each event. More information about the event annotation can be found in Kim et al (2008), http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=Event+Annotation and associated annotation guidelines: http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/release/GENIA_event_annotation_guidelines.pdf

Meta-knowledge annotation was carried out manually on top of the event annotations. More information about the meta-knowledge annotation can be found in Thompson et al. (2011). Also see: and http://www.nactem.ac.uk/meta-knowledge/.

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

Not applicable – this is a monolingual corpus.

4.4.4 Attributes and their values (if annotated)

An extract from a typical annotated file is as follows:

```
<sentence id="S2">We have examined the effect of
<term id="T6" lex="leukotriene_B4"
sem="Organic_compound_other">leukotriene B4</term>
(<term id="T7" lex="LTB4"
sem="Organic_compound_other">LTB4</term>), a
potent lipid <term id="T9"
lex="proinflammatory_mediator"
```

```
sem="Protein_family_or_group">proinflammatory
mediator</term>, on the expression of the <cons
id="T10" lex="(AND proto-oncogene_c-jun proto-
oncogene_c-fos)" sem="(AND DNA_domain_or_region
DNA_domain_or_region)"><frag id="F4">proto-
oncogenes</frag> <frag id="F5"><term id="A3"
sem="DNA_domain_or_region">c-jun</term></frag> and
<frag id="F6"><term id="A4"
sem="DNA_domain_or_region">c-
fos</term></frag></cons>.</sentence>

<event KT="Investigation" id="E7"
uncertainty="doubtful">
<type class="Regulation"/>
<theme idref="E9"/>
<cause idref="T7"/>
<clue>We have <clueKT>examined</clueKT> the
<clueType>effect</clueType>
<linkCause>of</linkCause> leukotriene B4 (LTB4), a
potent lipid proinflammatory mediator,
<linkTheme>on</linkTheme> the expression of the
proto-oncogenes c-jun and c-fos.</clue>
</event>
```

Each *sentence* tag contains term annotations. These are either indicated using *term* tags for simple, or additionally using *cons* and *frag* tags for more complex terms. An example of a complex term is a co-ordination in which two terms are contained, but ellipsis is involved. In the above example, this is exemplified by the phrase *proto-oncogenes c-jun and c-fos,* in which there are 2 terms, *proto-oncogene c-jun* and *proto-oncogene c-fos.* The *cons* tag surrounds the whole phrase, while *frag* is used to denote the individual parts of the terms, i.e., the common part of the two terms, *proto-oncogenes* and the unique parts, i.e., *c-jun* and *c-fos.* All of the above-mentioned tags have *id* attributes containing unique ids. *Term* and *cons* have the following additional attributes:

- *lex* attribute – the lexical representation of the term, which spaces replaced by underscores. In the case of the *cons* tag, this may include "AND" to show that two or more terms are contained within the text surrounded by the tag.
- *sem* – the semantic class assigned to the term from the GENIA term ontology. In the case of the *cons* tag, there may be multiple categories (one for each term within the enclosed phrase), indicated using "AND".

Tags of type *event* follow the *sentence* annotation and encode the events that have been annotated within the sentence. Each event has an *id* attribute to assign a unique id. Other attributes are optional, but can include the following (if non default values are assigned).

- *uncertainty* – one of the basic types of information relating to event interpretation annotated as part of the original event annotation. Can have the following values: *certain* (there is no doubt that the event took place; default value), *probable* (there is some level of speculation surrounding the event), *doubtful* (the event is under investigation). Note that the combination of meta-knowledge annotation dimensions (see below) is intended to provide more detailed information relating to event interpretation, and hence largely supersedes this attribute. However, it is retained for historical purposes.
- *assertion* – another one of the basic types of interpretation added as part of the original event annotation. Can have the following values: *non-exist* (the event is explicitly negated), *exist* (there is no explicit negation of the event; default value). This is somewhat similar to the *Polarity* meta-knowledge dimension. However, meta-knowledge is meant to encode more subtle differences, and so largely supersedes this attribute. However, it is retained for historical purposes.
- *KT [Knowledge Type]*, *CL [Certainty Level], Polarity, Manner, Source* – these correspond to the newly added meta-knowledge dimensions that add greater detail about the intended interpretation of events. Each has its own set of possible values. Further details can be found in Thompson et al. (2011) and the meta-knowledge annotation guidelines (http://www.nactem.ac.uk/meta-knowledge/Annotation_Guidelines.pdf)

The *clue* tag includes the text of the sentence in which the event is contained. Several clue expressions may be annotated, which are envisaged to help with the automatic recognition of events and associated meta-knowledge. The possible tags that can occur within the *clue* tag are as follows:

- *clueType* – The event trigger word or phrase
- *clueLoc* – the location in which the event took place
- *clueExperiment* – experimental techniques specified for the event.
- *clueTime* – corresponds to when the event happened or will happen.
- *linkCause* – used to indicate words that are used in the text link between and event and its CAUSE. They can be seen as words that "introduce" the CAUSE of the event, e.g. ***effect*** *of X on Y,* where *of* is the linkCause.
- *linkTheme* – used to indicate words used in the text to link the event and its THEME. They can be seen as words that introduce the THEME of the event, e.g. e.g. ***effect*** *of X on Y,* where *on* is the

linkTheme.

- *coRefCause* – annotated when the CAUSE of the event is an expression such as *it* or *this protein*, referring to a previously introduced (or coreferent) NE, either in the current sentence or in a previous sentence.
- *coRefTheme* – annotated when the THEME of the event contains an expression such as *it* or *this protein*, referring to a previously introduced (or coreferent) NE, either in the current sentence or in a previous sentence.
- *clueKT* – clue expression used to determine the chosen value of the *Knowlegde Type (KT)* meta-knowledge dimension.
- *clueCL* – clue expression used to determine the chosen value of the *Certainty Level* (CL) meta-knowledge dimension.
- *cluePolarity* - clue expression used to determine the chosen value of the *Polarity* meta-knowledge dimension.
- *clueManner* - clue expression used to determine the chosen value of the *Manner* meta-knowledge dimension.
- *clueSource* – clue expression used to determine the chosen value of the *Manner* meta-knowledge dimension.

4.5 Intended application of the corpus

The corpus is intended to allow to facilitate the advanced information semantic search systems in the biomedical domain, that allow events to be located using structured queries that can take into account semantic roles, NEs, etc. Systems trained to recognise meta-knowledge can allow extra sets of search criteria to be specified. This can allow, e.g., the isolation of new experimental knowledge to facilitate biomedical database curation, enable textual inference to detect entailments and contradictions, etc.

4.6 Reliability of the annotations (automatically/manually assigned) – if any

In terms of the meta-knowledge annotations, inter-annotator agreement rates in the range 0.84 – 0.92 Kappa were achieved, according to the different dimensions (Thompson et al., 2011)

5 RELEVANT REFERENCES AND OTHER INFORMATION

Kim, J-D., T. Ohta, Y. Tateisi, and J. Tsujii (2003). Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics,* 19:180i–182i.

Kim, J.-D., Ohta, T. and Tsujii, J.. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10

Thompson, P., Nawaz, R., McNaught, J. and Ananiadou, S. (2011). Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12:393