1. Meta-data for the ART Corpus

Title: ART Corpus
Data Type: Text
Data Sources: papers from journals of the Royal Society of Chemistry Publishing, in SciXML format.
Project: ART Project (http://www.aber.ac.uk/compsci/Research/bio/art/)
Recommended applications: information extraction, information retrieval, discourse analysis, natural language generation, text summarisation, machine learning for automatic annotation.
Languages: English.
Distribution: Download from (link).
Corpus Documentation: (link).

2. ART Corpus Documentation

Overall nature of the Corpus

Within the JISC funded ART project (University of Wales, Aberystwyth http://www.aber.ac.uk/compsci/Research/bio/art/) we developed a tool (SAPIENT)[4] to allow the annotation of scientific papers with core scientific concepts (e.g. 'Goal', 'Hypothesis', 'Experiment', 'Method', 'Result', 'Conclusion', 'Motivation', 'Observation'). These concepts constitute the CISP meta-data[1] and were verified through an on-line survey addressed to researchers. The CISP meta-data were accompanied by a set of guidelines for their implementation as an annotation scheme. We worked with chemistry experts, who used the guidelines and SAPIENT to create a corpus of 225 papers manually annotated with CISP concepts (ART Corpus > 1 million words, 35,040 sentences). These papers cover topics in physical chemistry and biochemistry and were provided by the Royal Society of Chemistry (RSC) Publishing.

The Corpus was developed primarily to add value to scientific papers, through semantic markup that would make it easier for natural language processing and semantic web applications to automatically extract information pertaining to core scientific concepts. The ART corpus can also be used as a training set for machine learning algorithms, in order to automate the annotation of papers with CISP meta-data. The sustainability of and the benefits obtained from annotating papers with CISP meta-data will be investigated by the JISC funded SAPIENT Automation (SAPIENTA) project.

Source Data: The source data consists of text in XML format, encoded in unicode (utf-8 character set). The XML schema used is a variant of SciXML[2], which can be provided upon request.
The differences between the ART Corpus XML and SciXML consist in the following:

- An <s> element has been added at the same level as the <S>, <EQN> and <EXAMPLE> elements. The latter elements can occur within a <P> element according to the SciXML schema. This <s> tag covers all kinds of sentences. That is, there is no distinction between sentences in the abstract (denoted as <A-S> in SciXML) and sentences in the main paper (denoted as <S> in SciXML) or sentences within equations (<EQ-S>) and examples (<EX-S>).
- The <s> element has an id (sid) and can include an <annotationART> element.
- The <annotationART> element has the attributes "type", "conceptID", "novelty" and "advantage". For more details please refer to the annotation guidelines[3].

Annotation: The goal of the annotation was to mark-up core scientific concepts in research papers. Papers from the domains of chemistry and biochemistry were chosen as a proof of principle approach. Annotation was performed by 20 chemistry experts, at PhD or postdoctorate level with excellent knowledge of English. The annotators selected were given an annotation package consisting of a set of guidelines[3] for annotating papers with CISP, the SAPIENT system[4] and its manual, as well as an example paper which had already been annotated. Most of this material is available for download from: http://www.aber.ac.uk/compsci/Research/bio/art/sapient.
The annotation guidelines are available upon request.

Work with annotators was conducted in three phases over a period of six months.

In phase I (training phase) all 20 annotators were sent the same four papers to annotate using SAPIENT and the annotation guidelines, in order to familiarise themselves with the process. Individual annotators' results were analysed meticulously at this stage and were used to improve the guidelines.

For Stage II, (evaluation phase) the aim was to evaluate both the annotators and the guidelines. A preliminary evaluation of the experts' agreement was conducted based on a sample of 41 papers (5,000 sentences) which were annotated by 16 experts, divided in non-overlapping groups of 3 experts. The results show significant agreement between annotators, given the difficulty of the task (an average kappa co-efficient of 0.55 per group).

The 9 experts from phase II who had the highest average inter-annotator agreement were selected for phase III. The latter constitutes the actual creation of the ART Corpus, through the annotation of 225 papers.

Distribution: The ART corpus is available as a 2.2 MB tar.gz file which expands to 12 MB. It consists of 225 papers (> 1 million words, 35,040 sentences). The corpus is available as a collection of 225 .xml files, where each file corresponds to a separate paper whose sentences have been annotated individually with core scientific concepts. The papers have been arranged into 9 folders, corresponding to each of the 9 annotators. These papers can be processed individually, per folder or as a batch by any script for handling XML.

One can display papers individually by using the SAPIENT software[4], which was used for creating the original annotations. For instructions on how to use SAPIENT to display the software please refer to SAPIENT_FAQ.txt (both can be downloaded from: http://www.aber.ac.uk/compsci/Research/bio/art/sapient.)

For any requests/details regarding the corpus please contact Dr Maria Liakata (mal@aber.ac.uk).

References

[1] Soldatova L. and Liakata M. (2007). An ontology methodology and CISP - the proposed Core Information about Scientific Papers. JISC Project Report. http://ie-repositor y.jisc.ac.uk/137/, 2007.

[2] CJ Rupp, Ann Copestake, Simone Teufel and Ben Waldron. (2006). Flexible Interfaces in the Application of Language Technology to an eScience Corpus. Proceedings of the UK e-Science Programme All Hands Meeting 2006 (AHM2006), Nottingham, UK.

[3] Liakata M. and Soldatova, L. (2008). Guidelines for the annotation of General Scientific Concepts. JISC Project Report. http://ie-repository.jisc.ac.uk/.

[4] Liakata Maria, Q Claire and Soldatova Larisa N. (2009). Semantic Annotation of Papers: Interface & Enrichment Tool (SAPIENT). *To appear* in Proceedings of BioNLP at NAACL, Boulder, Colorado.