SAPIENT Automation project

Dr Maria Liakata
Leverhulme Trust Early Career fellow
Department of Computer Science, Aberystwyth University
Visitor at EBI, Cambridge
mal@aber.ac.uk

25 May 2010, London

SAPIENT Automation Project

- **Goal:** Help researchers process scientific papers faster and get the information they are interested in out of them.
- **How:** By automating the recognition of core scientific concepts (CoreSCs) in papers and evaluating them in terms of usefulness in user based summaries.
- **CoreSCs:** Hypothesis, Motivation, Goal, Object, Background, Method, Experiment, Model, Observation, Result, Conclusion
- **Envisaged Outcome:** A system for automatic recognition of CoreSCs in papers, a method for creating extractive user based summaries and an evaluation of the usefulness of CoresCs.

Scientific papers

- Plethora of scientific papers produced at an increasing rate: a challenge for scientists; researchers, reviewers, curators.
- A challenge for automatic methods to extract and summarise information from scientific papers
- State of the art deals mainly with abstracts (especially in the Biosciences)
- Initiatives such as Sciborg and Flyslip (Cambridge) to work with full papers, NaCTeM has started annotation of full papers
- Issues: format, content, structure

The ART project

- Has produced a set of meta-data CISP and an implementation of it a three layered annotation scheme (CoreSC) for annotating full scientific papers.
- CoreSC and CISP views a paper as a humanly readable version of a scientific investigation and seeks to discover this in texts
- ART produced annotation guidelines for CoreSC and a corpus of 265 papers in physical chemistry and biochemistry annotated by 16 experts at the sentence level with CoreSCs (40,000 sentences, $>$ 1 million words)
- Papers are in SciXML
- ART also produced an annotation tool SAPIENT, to aid experts in the manual annotation of CoreSCs

SAPIENT: An interface for semantic annotation

- **SAPIENT** is a web-based tool so that it can be incorporated in editing workflows, platform independence.
- Developed for **sentence by sentence** annotation of full papers in XML.
- Incorporates an XML aware sentence splitter, **SSSplit** which works with all XML schemas.
- SAPIENT can be used to annotate papers with CoreSC concepts and incorporates OSCAR3.
- Can also be used with other sentence based annotation schemes.
- Suitable for manual annotation, currently being automated.
- SAPIENT and SSSplit developed at Aberystwyth http://www.aber.ac.uk/en/cs/research/cb/projects/sapienta/software/.

SAPIENT Important Features

- Allows sentence based **annotation at multiple levels**. Can specify properties/attributes.
- SAPIENT generates both sentence identifiers and annotation identifiers (**concept IDs**).
- Concept IDs **link sentences** pertaining to the **same instances** of an annotation **concept** forming **zones of interest**.
- Automated noun phrase based annotation from existing ontologies available through **OSCAR3**
- Has been used by 16 experts for annotating **270 papers**.
- Can accept Pubmed Central papers and various **XML schemas**.
- **Plans** for allowing multiple annotations per sentence.

Core Scientific Concepts (CoreSC)

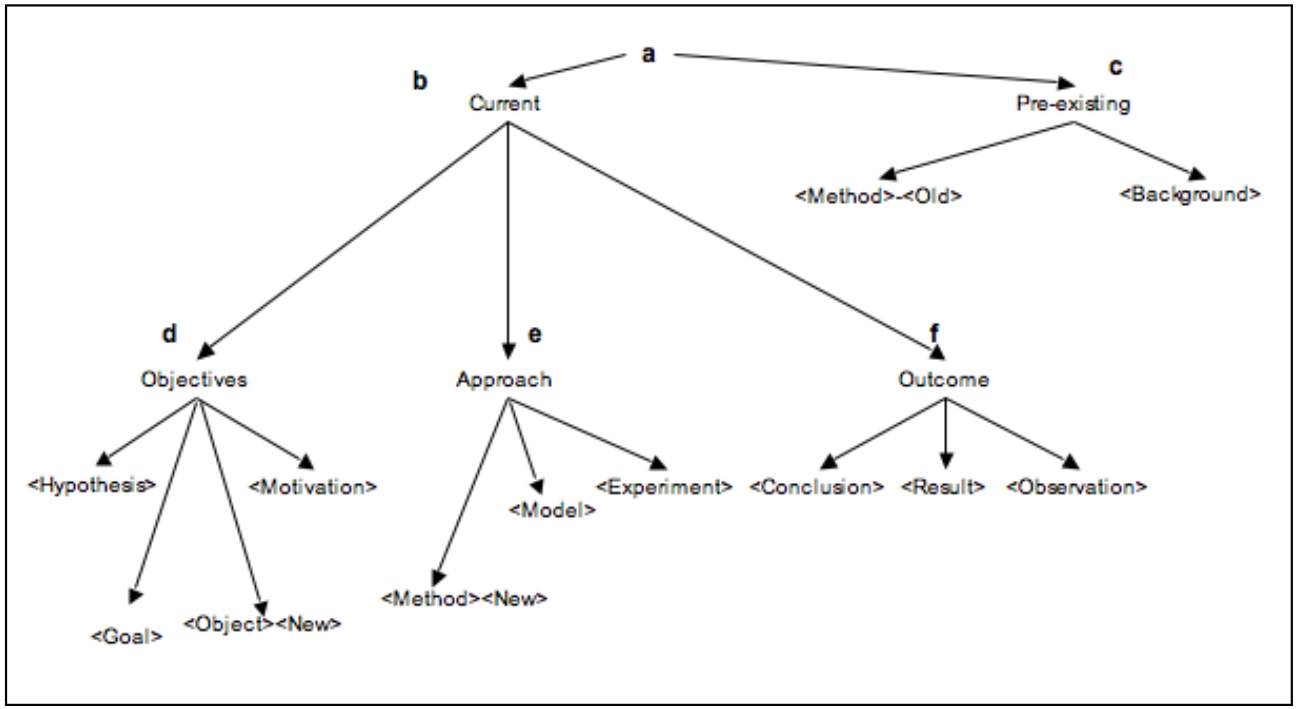

Figure: A taxonomic representation of CoreSCs

CoreSC annotation scheme

- Three layered annotation scheme:
  - 1st layer: Core Scientific Concepts ("Motivation", "Hypothesist", etc.)
  - 2nd layer: Novelty (New/Old) and Advantage (advantage/disadvantage)
  - 3rd layer: Identifiers, linking sentences referring to same instance of a CoreSC

Example CoreSC annotation

in a sample without the need for separation. </annotationART></s><s sid="24"><annotationART atype="GSC" type="Obj" conceptID="Obj1" novelty="New" advantage="Yes">In <ne id="o44" surface="addition" type="ONT" rightPunct=",">addition</ne>, it is an extremely sensitive technique with single <ne id="o45" surface="molecule" type="ONT">molecule</ne> detection reported<REF TYPE="P" text="6,7" ID="cit6 cit7">6,7</REF>. </annotationART></s><s sid="25"><annotationART atype="GSC" type="Met" conceptID="Met5" novelty="New" advantage="None">The work reported here was carried out using <ne id="o46" surface="citrate" type="CM" confidence="0.9123527745636536">citrate</ne> reduced <ne id="o47" surface="silver" type="CM" confidence="0.9883660308368389" Element="Ag" ontIDs="CHEBI:33966">silver</ne> nanoparticles since, by careful control over the aggregation and experimental conditions, quantitative and reproducible results can be obtained<REF TYPE="P" text="8,9" ID="cit8 cit9">8,9</REF>. </annotationART></s><s sid="26"><annotationART atype="GSC" type="Obj" conceptID="Obj1" novelty="New" advantage="Yes">A major benefit of <ne id="o49" surface="SERRS" type="CM" confidence="0.21115177068959173">SERRS</ne> is that <ne id="o50" surface="fluorescence" type="ONT">fluorescence</ne> is efficiently quenched by the <ne id="o51" surface="metal" type="ONT">metal</ne> surface allowing a large range of coloured <ne id="o52" surface="molecules" type="ONT" rightPunct=",">molecules</ne>, including standard fluorophores, to be used as SERRS labels.</annotationART></s></P

Index | Refresh | Auto Annotate | Clear Auto Annotations | Clear Own Annotations | Save | Help

ectra to be obtained,

Background Conclusion Experiment Goal Hypothesis Method Model Motivation Object Observation Results

24 In addition, it is an extremely sensitive technique with single molecule detection reported[6,7].

Object | Advantage | Obj1

25 The work reported here was carried out using citrate reduced silver nanoparticles since, by careful control over the aggregation and experimental conditions, quantitative and reproducible results can be obtained[8,9].

Method | New | Met5

26 A major benefit of SERRS is that fluorescence is efficiently quenched by the metal surface allowing a large range of coloured molecules, including standard fluorophores, to be used as SERRS labels.

Object | Advantage | Obj1

Project objectives and issues to explore

- Evaluate existing meta-data in ART corpus in terms of statistical measures
- Automate the recognition of CoreSCs in full papers using machine learning and integrate with current annotation interface
- Evaluate the methods and meta-data on a testing set from the ART/CoreSC corpus and on a new set of papers
- Use automatically generated CoreSCs to create user based extractive summaries
- Evaluate the usefulness of the CoreSCs in terms of user experience (extractive summaries but also other purposes, e.g. teaching materials)
- Evaluate CoreSCs in terms of ease/accuracy in recognition using machine learning and compare against other annotation schemes
- Smooth integration of the automated methods into the SAPIENTA (SAPIENT Automated) software.
- Investigation into new types of queries over the meta-data, which can add to the functionality of SAPIENTA.
- Disseminate SAPIENTA and the automatic meta-data to the research community and target particular users (reviewers, editors, authors)

What has been achieved so far

- **Evaluate existing meta-data in ART corpus in terms of statistical measures**
- **Automate the recognition of CoreSCs in full papers using machine learning**
- *Integrate automated methods with current annotation interface*
- Evaluate the methods and meta-data on a testing set from the ART/CoreSC corpus and on a new set of papers
- *Use automatically generated CoreSCs to create user based extractive summaries*
- Evaluate the usefulness of the CoreSCs in terms of user experience (extractive summaries but also other purposes, e.g. teaching materials)
- **Evaluate CoreSCs in terms of ease/accuracy in recognition using machine learning and compare against other annotation schemes**
- *Smooth integration of the automated methods into the SAPIENTA (SAPIENT Automated) software.*
- *Investigation into new types of queries over the meta-data, which can add to the functionality of SAPIENTA.*
- Disseminate SAPIENTA and the automatic meta-data to the research community and target particular users (reviewers, editors, authors)

KEY: **Bold** is done, *Italic* is future and Regular is current

Evaluation in terms of statistical measures

- Considered corpus as two phases of development
- Looked carefully at kappa inter-annotator agreement in phase I
- Used the results to organise papers according to annotation quality into two tiers

Automating the recognition of core scientific concepts in full papers

- Used ART/CoreSC corpus as training data.
- Selected appropriate features (location of sentence, length, n-grams: over 50,000 features)
- Considered different types of cross-validation (annotator, pseudo-random folds, test-train on subsets of the tiers)
- Text classification using SVM. Obtained preliminary results of 40-50% F-measure
- Some categories a lot easier to recognise than others. Hypothesis, Model difficult ones
- **Currently** improving features, adding POS and GR information
- Problem: No parser has been evaluated on full biomedical papers so missing over 10% sentences
- Also try CRF machine learning algorithm

Comparison with other schemes and user evaluation of CoreSCs

- Performed comparison of CoreSC and AZ-II: Discovered that are complementary schemes (LREC 2010 paper, in collaboration with Simone Teufel, Advaith Siddharthan and Colin Batchelor)
- Recognised core scientific concepts in 1000 abstracts (in collaboration with the CRAB project, University of Cambridge)
- Compared against two more coarse-grained schemes in terms of ease of recognition using ML (variant of AZ and a scheme for abstracts)
- 80% F-score for CoreSCs in abstracts, lower than for the other two but a lot more fine grained scheme
- **Currently:** User tests in terms of the extent to which CoreSC categories help Cancer Risk Assessment (CRA)
- Annotation of full papers for CRA using CoreSCs

Extractive summaries and other plans

- Generate user-based extractive summaries from the CoreSC meta-data after consultation with experts
- Combination with other annotation schemes (NaCTeM, Teufel) and evaluation. Annotate same papers as NaCTeM, map annotation schemes
- Linking to information from existing ontologies
- Explore negation in the papers

Exploring other uses for CoreSCs

- Use CoreSC papers for IE task, with and without the annotations. Any improvements?
- Intelligent querying of the papers:
  - Parse the ART/CoreSC corpus
  - Obtain logical forms for targetted sections
  - Use reasoning for Question Answering
- Learning/populating a domain ontology: provides guided sections for finding relations and class instances.
- Uses in publishers editing workflow?
- Combine above components to form system of automatic reviewing of papers

Other short projects of interest to us

- Sentiment analysis to be combined with CoreSC categories
- Comparison of abstracts and full papers in terms of CoreSC distributions.
  To what extent are abstracts representative summaries?
- Incorporate better system for NE/term recognition
- How to learn identifiers for finding same instances of a particular concept.
  (i.e. How do you tell where Obs1 ends and Obs2 begins?)

What to look out for in the future

- Make sure necessary resources are available. We did not have enough computational power.
- Anything that involves computation needs at least twice as long as originally envisaged.
- Engagement with users very beneficial
- Help needed with dissemination and engagement with users

Stake holders and beneficiaries

- **Readers of scientific papers will benefit from ease of access to information content**
- Knowledge experts, scientists, researchers, editors, reviewers, curators
- **Potential significant impact to publishers' workflow**
- Students, learners, authors of papers. Examples of Methods, Hypotheses, Conclusions clearly outlined and highlighted
- Professor Ngo to use in his teaching materials
- Experts have said that CoreSCs help them write papers in a better way
- Potential advances in summarisation and textual inference
- Evaluation of parsing in full biomedical papers
- Dissemination and interaction with related initiatives will be of key importance

THANK YOU for listening!!
Any Questions?