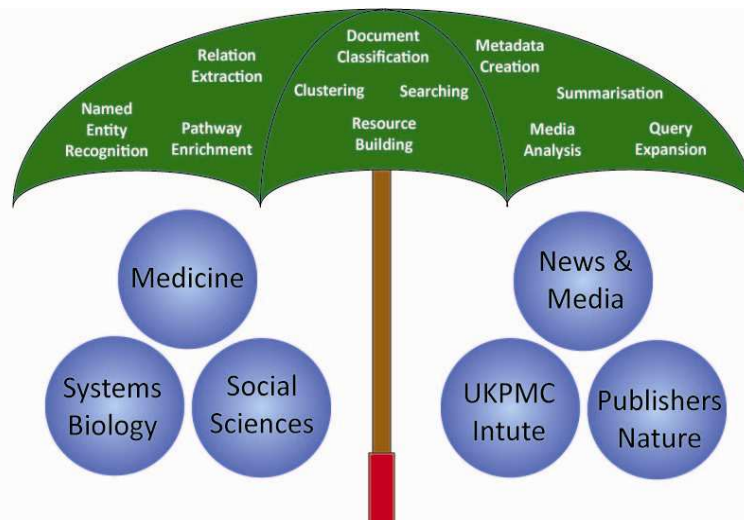


Welcome to an extended edition of the NaCTeM newsletter. In this issue we bring together some of the key events and activities so far in the second phase of NaCTeM. Building upon the core technologies developed for biology during phase I. We are now broadening our range of domains to also cover health and medical sciences, social sciences as well as the more general area of scholarly communication.

More information about the centre can be found on our website at <http://www.nactem.ac.uk> or via a selection of surveys and briefing papers available at:

- Text Mining Briefing Paper - <http://jisc.ac.uk/media/documents/publications/bptextminingv2.pdf>
- Introduction to NaCTeM - <http://jisc.ac.uk/media/documents/publications/bpnationalcentrefortextminingv1.pdf>
- Vision for the future - <http://www.ariadne.ac.uk/issue53/ananiadou/>



Updated MEDIE Makes Fact Finding Easier

Service: MEDIE - Enhanced semantic retrieval of relations and events

A new version of the MEDIE semantic search engine is now available. MEDIE retrieves biomedical relations and events from MEDLINE™, based on deep analysis using the Enju parser and semantic search with biomedical entities, such as genes and diseases. This deep analysis allows MEDIE to extract relations from abstracts, compensating for variability in word choice and word order.

The user specifies relations to find, such as: "*What activates p53?*" or "*What does dystrophin trigger?*" and is then given a list of documents showing the direct occurrence of the supporting evidence.

Further features of MEDIE include:

- Searching the most up-to-date abstracts: the system is updated daily to include new abstracts added to MEDLINE™.
- Searching for relations in specific sections of the abstract, such as title, objective, method, result, or conclusion.
- Optional use of ontology terms in the search.
- Optional specification of journal and/or author names to narrow the search.

For further details regarding this tool follow the links for MEDIE from the NaCTeM Services page <http://www.nactem.ac.uk/services.php>

BioLexicon
Coming Soon

visit
<http://www.nactem.ac.uk/biolexicon/>
for more information

You Came, You Saw, You Conquered – U-Compare(d)

Infrastructure: Open framework for evaluation and interoperability of text mining tools

Text mining tools come in a variety of forms, even when they support the same task. The choice of a single tool from the range can therefore be a daunting task for an expert, who can appreciate the subtle differences between the architectures and training sets, let alone a novice just looking to process their data. Now, researchers in a collaboration between the University of Tokyo, NaCTeM and the University of Colorado Health Science Center have developed a service that can help, resulting in its latest offering called U-Compare.

Underlying the framework is the Unstructured Information management Architecture ([UIMA](#)) system, which allows for text mining tools to interact through a shared type system. This enables tools from developers all around the world to work together and only requires a simple wrapper that transforms the input and output to this shared type system. Originally this was used to build up chains of tools to act in a workflow to produce complex analysis networks. With U-Compare this has now been extended to include evaluation tools that enable users to see the effects on output of the choice of tools, in order to find the best methods required to solve an individual task. U-Compare can also be plugged into popular workflows like [Taverna](#).

This marks a significant step in making text mining tools more flexible, comparable and approachable to new users. We look forward to the community supporting this infrastructure.

Find out more at <http://u-compare.org/>



“A rose by any other name would smell as sweet”

Resource: Linguistic Resources for the Biomedical Community

The BioLexicon is a collective achievement by EMBL-EBI, CNR-ILC, and the University of Manchester in the EC BOOTStrep Project. It is an integrated lexical-terminological resource containing around 2.2 million lexical entries suitable for biological text mining.

Biological terminology is a frequent cause of analysis errors when processing biological literature. For example, "retro-regulate" is a terminological verb often used in molecular biology but it is not included in conventional dictionaries. This linguistic resource will be a great assistance for those who need to process text in the biology domain. The BioLexicon will become available in the very near future from ELRA (<http://www.elra.info/>). More information on the BioLexicon can be found at <http://www.nactem.ac.uk/biolexicon/>

Healthy Results at User Day

Event: Text Mining Workshop for Health – 26/03/09

The latest of our domain focussed workshops was recently held in Manchester, highlighting the benefits and practical uses of text mining in everyday activities. This led to an excellent start for the upcoming season of events with an extra session being added to cater to popular demand by the growing audience, eager to learn more about the techniques that go beyond simple search.

The session began with an overview of what text mining is and how it can benefit users, followed by a discussion of specific examples where text mining has provided new and important hypotheses within the world of health. After a break out into smaller groups the 92 attendees were presented with the core technologies, which were discussed and backed up by a demonstration and hands on session of relevant NaCTeM services; [Kleio](#), [Facta](#), [MEDIE](#) and [ASSERT](#). Finally, a chance to discuss any burning issues and specific challenges over coffee was made available at each session with considerable response by our guests.

Highlights from their feedback include:

MEDIE is fast (72%) and directly applicable to their work (75%)

Facta is fast (70%) and easy to use (73%)

Kleio is fast (79%) and easy to use (73%)

ASSERT is fast (80%) and directly applicable to their work (83%)

Healthy Results at User Day cont...

User Comments:

On ASSERT: "Great as it is - if improved -> MINDBOGGLING"

On KLEIO: "Would like to search across full articles"

On FACTA: "Indirect associations would be useful"

On MEDIE: "Store searches"

With respect to these comments, we are currently in the final stages of testing our indirect relationships version of Facta, check back on our website soon for more information on updates to this and other NaCTeM tools. Further to this as part of the UKPubMedCentral project, some of the biomedical tools will be updated to include full text data from the PubMedCentral resource. This will offer a richer analysis and better results through access to the core content, not just the summaries that often make up an abstract. Other comments about functionalities are carefully being analysed to prioritise these for new functionality for the services. NaCTeM would like to thank all of those who attended for their feedback on our tools and would also like to offer special appreciation to the Faculty of Medical and Human Sciences at the University of Manchester for their support in the event.

Political Review

Update: TerMine analysis of Political Speeches

Have you ever had to read through a long document with a poor abstract, when all you wanted to know was is this document worth reading? See how you can use the TerMine service to move beyond the increasingly popular tag clouds. By identifying multi-word technical terms the Termine service is able to detect core concepts being discussed and rank them by importance, giving you a fast and effective list of topics weighted by their strength.

See how this works on our new TerMine examples page <http://www.nactem.ac.uk/software/termine/>. Latest examples include speeches by President Barack Obama and Prime Minister Gordon Brown as well as Alistair Darling's budget speech. The same tools are being used for document analysis, identifying core topics and using these results to improve existing search services. Try it for yourself.

Scalable Solutions for Repositories

Projects: Text mining services for the analysis of large-scale repository content

Growing numbers of repositories, both institutional and discipline specific, alongside greater rates of deposit are leading towards a long awaited resource granting access to the UK research literature, teaching materials and similar documentation. However, care needs to be taken during these early stages to ensure we do not end up with another version of the World Wide Web and the well discussed problems of information overload and overlook. Text mining technologies can assist across the whole repository lifecycle from deposit to preservation, but only if the correct protocols are put in place to support such activity.

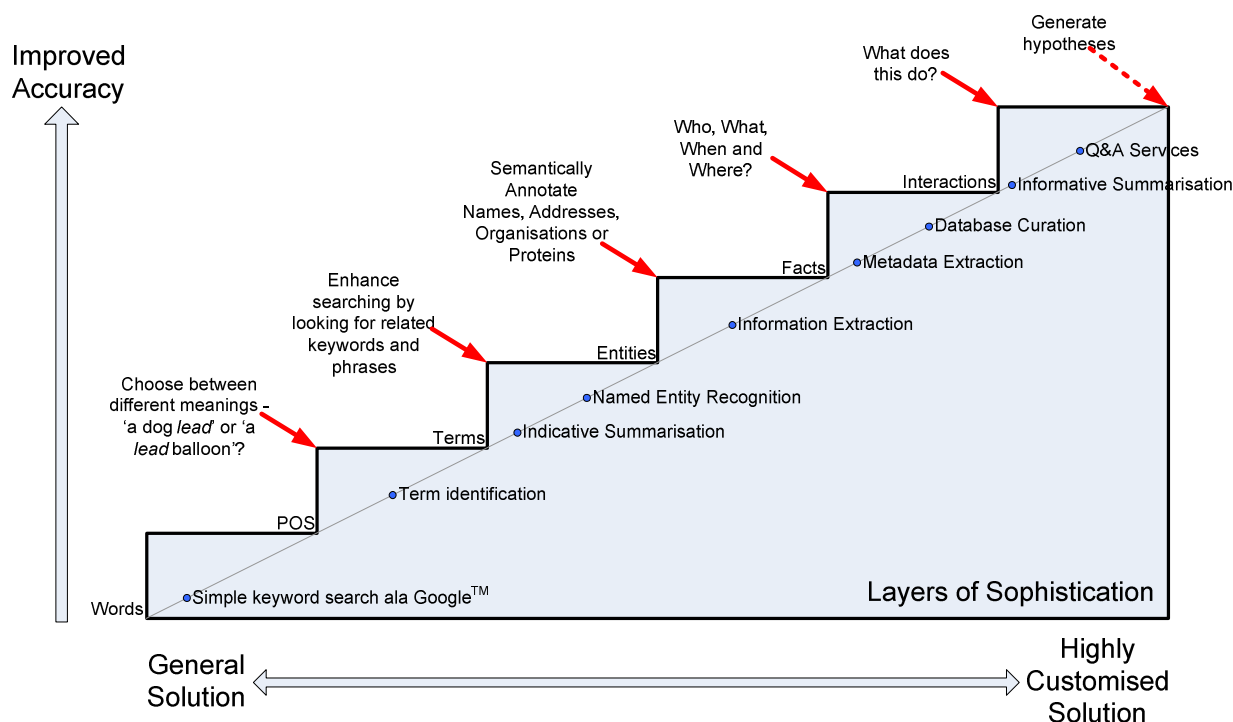
At NaCTeM we have been working towards this area for a number of years and have developed mature software capable of providing the required scalability, reliability and flexibility to support these activities, resulting in two main exemplars. The [Intute Repository Search](#) project has shown that automatic metadata creation/enhancement is possible across very general collections of texts; here we have used NaCTeM tools to mine the resources themselves, not just the Dublin Core descriptions, to great effect. Further to this the current [UKPubMedCentral](#) project, involving NaCTeM, is showing how this can be extended to support domain specific repositories, with tools highly customised to assist biomedical researchers in their everyday information analysis tasks.

Steps to Success

Feature: The complexity of text mining – layers of solution

Text mining solutions cover a wide range of techniques from basic pattern matching around words to in-depth natural language processing for detecting the syntactic relationships in a sentence. Each is based on different levels of knowledge about the content, usually represented as annotations. As each layer of annotation is added, the level of sophistication is raised allowing for greater opportunities to exploit the information in new ways. However to construct these annotations requires more complex tools which need to be customized and trained on domain specific examples. This trade off between generalization of the tools and sophistication of the analysis allows for a wide variety of solutions for the user, but comes at a cost of additional work and resources being required.

The NaCTeM exemplars show what is possible with the core tools, but their flexibility depends on where they are based in the layers of sophistication. To support other scenarios, or similar scenarios in different domains, further services are required which rely on additional tool customization. This customization is necessary to adapt tools to handle the differences in writing style and content.



For example it is useful to compare the TerMine service with MEDIE. TerMine is near the left of the diagram and as such is easily ported between different domains as can be seen in our ASSERT, ASSIST, BBC, Intute Repository Search and other projects. In contrast MEDIE focuses on the deep syntactic analysis using machine learning methods trained on large volumes of annotated texts. This training set is in itself a valuable resource, and can be very costly to build up demanding domain experts' time to manually annotate items of interest to improve the fact extraction.

As you can see, different solutions will suit different tasks and different levels of sophistication. The key message is that customization is required to target specific user groups but it is underpinned by our generic text mining software which can be adapted according to user requirements. The choice for the user then becomes what level of sophistication will give me appropriate results, within a given schedule or available resource.

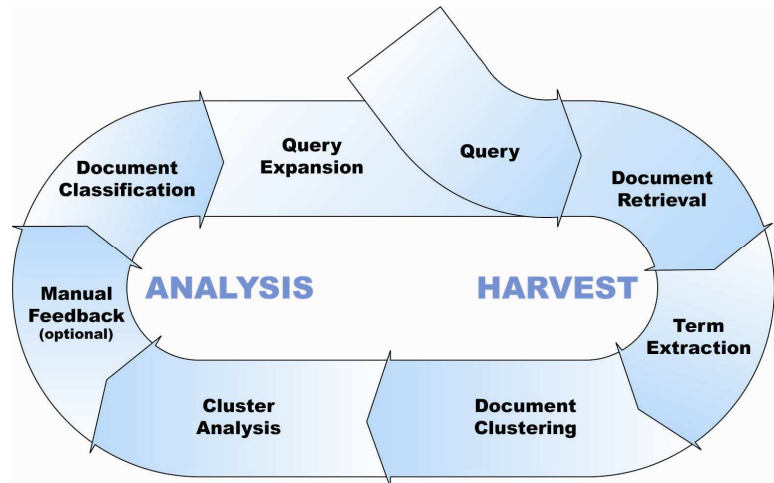
Systematic Reviewing Assisted by Text Mining

Service: ASSERT Project Demonstrator Now Available

A new web demonstrator has been released allowing users to explore the core technology used in the JISC funded Automatic Summarisation for Systematic Reviews using Text Mining (ASSERT) project. The complete tool aims to assist users in constructing systematic reviews and has been implemented in close collaboration with experts at the EPPI-Centre in order to complement manual workflows.

Further features of ASSERT include:

- **Topic Detection** to help you discover and browse the key concepts within your results
- **Query Expansion** to help find information you may have otherwise overlooked
- **Document Clustering** to help you break your results into manageable sets
- **Multi-Document Summarisation** to help make sense of all the content and find important sentences in your results.



Further details about the research in this project and the techniques used can be found at <http://www.nactem.ac.uk/assert> along with links to the demonstrator and a video walkthrough of the system.

High-throughput Text Mining with HPC

Infrastructure: Early Stage Experiments with Parallelising NaCTeM Tools

With nearly 5000 scientific articles published daily just within the biomedical domains it is easy to see why text mining is becoming an essential technology in this information age. By adding to this collection: additional domains, unpublished articles, news feeds, grey literature and a plethora of other sources and you can see that, at the current rate of growth, even cutting edge tools may begin to struggle with the load, particularly when you take into account the backlog of literature already available. For example the MEDLINE™ database contains over 18 million articles, usually only abstracts. This in itself is a challenge but when you consider the analysis of full text articles, often containing around 20 times more content, the true scale of the challenge can be observed.

With this in mind, NaCTeM researchers have recently carried out a detailed trial on a number of computing environments, including the Distributed European Infrastructure for Supercomputing Applications ([DEISA](#)), using core NaCTeM tools and document collections containing MEDLINE™ abstracts, and full text articles from the PubMedCentral collection. Initial results have shown that the parallelising process clearly reduces the time needed for analysis to a more manageable figure. In the case shown above, a user could have their results in less than 16 hours rather than having to wait more than 74 days. Though this experiment has been successful, further challenges remain before these tools can be made available such as data storage and transfer issues, but future work will investigate these in more detail and we will make our results available in due course. Further information on these results and technical details will be made available in a forthcoming report on the NaCTeM website.

Comments from our Communities

“Text mining has exciting applications for medicine. Conventional sifting of information can take weeks, and exciting new connections could potentially be missed. Medical research is also increasing interdisciplinary, including biology, chemistry, economics and other sciences. Being able to access information from other fields is a tremendous benefit and can help generate new ideas. Access to NaCTeM will be a real boost for our research teams, and a great incentive for new recruits.”

Professor Phil Baker, Director BRC

“[Sophia Ananiadou] discussed some of NaCTeM's flagship tools like MEDIE, FACTA and KLEIO - it does look like they're starting to take all the pain out of text mining, by doing the difficult bits for us, so we can use the results to do actual mining.”

Andrew B. Clegg – Blog post at biotext.org.uk [\[link\]](#)

“Over the last couple of years, scientists at Pfizer's UK research site in Sandwich have been making use of the text mining tools and services developed by NaCTeM. One such tool, which has proven to be valuable, is TerMine, an automatic multi-word term recognition tool that has been used at Pfizer to enrich the labour-intensive process of building dictionaries used for text mining. [...]

Pfizer and NaCTeM have also been collaborating on a project called DECA (Disease Extraction with Concept Association) to extract associations between concepts in the biomedical domain such as diseases and symptoms from collections of biomedical texts (e.g. Medline). The aim of this project is to combine the strengths of the NaCTeM text mining tools, Kleio and FACTA to create an efficient search for associations between biomedical concepts. Also, a considerable amount of research is being applied to the challenge of lexical disambiguation of the biomedical terms. Pfizer values highly the world-class quality of the linguistic and semantic extraction skills and methodologies being developed and practised at NaCTeM which is located in the highly appropriate setting of the Manchester Centre for Integrative Systems Biology.”

Ian Harrow, Senior Principal Scientist, Pfizer

“NaCTeM has engaged closely with users in systems biology to understand their needs and to provide cutting edge text mining services. Researchers in systems biology need integrated approaches to generate hypotheses and the use of text mining technology is a must for facilitating scientific discovery given the amount of textual data generated daily. NaCTeM has tapped into this potential with great success. One of the most impressive outcomes of the work of NaCTeM are the systems MEDIE and FACTA. Such semantically based tools are important for the discovery of new knowledge in biology.”

Professor Douglas Kell, Research Chair in Bioanalytical Science at the University of Manchester

“Sophia Ananiadou from NaCTeM explained the work her group has done using text mining techniques on Medline abstracts. This is the third time I've heard her talk about this, and it gets more interesting each time. Her aim is to enrich the literature by automatically creating semantic metadata, and thereby to make “undiscovered science” accessible. The MEDIE system is the most vivid example she showed, allowing you to construct a query in the form “subject – verb – object”. For instance, you can ask “what does p53 activate” by searching for `subject=p53, verb=activate.`”

Frank Norman – Trading Knowledge Blog [\[link\]](#)

Awards and New Projects

NaCTeM works side-by-side with the Text Mining Research Group at the University of Manchester to enable cutting edge technologies to develop into our services and to allow close collaboration with expert text miners.

JISC, CheTA - Chemistry using Text Annotations , University of Cambridge, working in partnership with the Royal Society of Chemistry (RSC) and Thomson Reuters	2009-2010
BBSRC BB/G013160/1, Automated Biological Event Extraction from the Literature for Drug Discovery , with AstraZeneca	2009-2012
BBSRC, Japan Partnering Award	2009-2012
Nature Publishing Group, commercial development project.	2009
Wellcome Trust, UKPubMedCentral , with MIMAS and Faculty of Life Sciences at the University of Manchester and the European Bioinformatics Institute	2008-2011
BBSRC BB/F006012/1, From data to knowledge – the ONDEX System for integrating Life Sciences data sources , led by Rothamsted Research.	2008-2011
JISC, ASSIST, Text mining for Social Sciences , in collaboration with NCeSS http://www.ncess.ac.uk/ and EPPI http://eppi.ioe.ac.uk/cms/	2008-2009
Pfizer Industrial Sponsorship, Disease named entity recognition and association mining	2008-2009
JISC, Metadata creation for Intute Repository Search , with MIMAS and UKOLN	2008-2009
Sophia Ananiadou was awarded the prestigious IBM UIMA Innovation Award [link] for the 3 rd time.	

More details are available on the NaCTeM website – http://www.nactem.ac.uk	
Services	http://www.nactem.ac.uk/services.php
Events	http://www.nactem.ac.uk/events.php
Publications	http://www.nactem.ac.uk/publications.php
News	http://www.nactem.ac.uk/news.php

If you would like to be involved in any of the events hosted by NaCTeM, please contact:
NaCTeM, Manchester Interdisciplinary Biocentre, 131 Princess Street., M1 7DN, Manchester
Tel: +44 (0) 161 306 3096 Fax: +44 (0) 161 306 3099

This NaCTeM Newsletter was edited by Brian Rea
Email: brian.rea@manchester.ac.uk
The deadline for the next issue is June 17th 2009