

What is Intelligent Text Processing?

Human language presents itself to us in both spoken and written form. Intelligent Text Processing (ITP) is concerned with adding value through the processing of written text, whether that text is in documents on your hard drive or on a web site. Typical ITP tasks are:

- text retrieval, where the goal is to locate a document that contains content the user is interested in, with results typically being found on the basis of a collection of keywords provided by the user; this kind of technology underlies the now ubiquitous Web search engines;
- text classification, where the aim is to identify which of a number of predefined categories a particular document belongs to: this is a technology with wide and varied application, already hiding just under the covers on many people's desktops where it is used to automatically filter out spam and junk mail;
- text summarisation, where the goal is to produce a shortened version of a document that captures the salient points, thus removing the need for the user to read the full original;
- question-answering systems, where, given a natural language question, the aim is to retrieve a likely answer to that question given a large body of text (such as the web, or a corporate intranet document store);
- information extraction, which identifies and extracts key elements of information from a given set of documents: for example, we might extract from a set of press releases key information about new products that have been released; and
- machine translation, which takes a document in one human language and translates it into another human language.

All of these applications have found their way out of the research laboratories and are producing commercially useful results, although the degree of activity in Australia and New Zealand is at present far less than in the US and Europe. We'll look more closely at many of these areas of technology in upcoming issues of LT Update.

Appen wins two awards for export activity

Appen, one of LT Update's supporting subscribers, has won two awards at the Premier's NSW Exporter of the Year Awards 2002. In its first ever entry in the competition, Appen received a 'Highly Commended' award in the Information and Communications Technology category; and the company picked up a second award when Marketing Manager, Christoph Vonwiller, was announced as the 'Premier's NSW Young Exporter of the Year 2002'.



Upcoming Events

National

- **Voice World 2003:** 26-27 February 2003, Sydney Convention and Exhibition Centre, Darling Harbour, Sydney. http://www.ccworldnet.com/2003/voice_AU/.
- **Joint International Conference: 4th ICCS International Conference on Cognitive Science and 7th ASCS Australasian Society for Cognitive Science Conference:** 13-17 July 2003, University of New South Wales, Sydney. <http://www.cogsci.unsw.edu.au>.

International

- **EACL 2003, 10th Conference of the European Chapter of the Association for Computational Linguistics:** 12-17 April 2003. Budapest, Hungary. <http://www.conferences.hu/EACL03/>.
- **ACL 2003, 41st Annual Meeting of the Association for Computational Linguistics:** 7-12 July 2003, Sapporo, Japan. <http://www.ec-inc.co.jp/ACL2003/>.
- **Corpus Linguistics 2003:** 28 March - 1 April 2003, Lancaster University, UK. <http://www.comp.lancs.ac.uk/ucrel/cl2003/>.
- **INTERACT 2003, Ninth IFIP TC13 International Conference on Human-Computer Interaction:** 1-5 September 2003. Zurich, Switzerland. <http://www.interact2003.org>.
- **ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems:** 28-31 August 2003, Chateau-d'Oex-Vaud, Switzerland. <http://www.speech.kth.se/error/>.
- **UM 2003, 9th International Conference on User Modeling:** 22-26 June 2003. University of Pittsburgh Conference Center, Johnstown, Pennsylvania, USA. <http://www2.sis.pitt.edu/~um2003/>.

Language Technology goes Polynesian

Most work in language technology has focussed on English and the languages of the dominant European and Asian countries. Bucking this trend, researchers at Otago University in New Zealand are developing a system that allows users to converse with a computer in either English or Maori. The system has a sophisticated set of dialogue-handling techniques that support a computer-aided language learning system aimed at helping learners of Maori practice some simple conversational skills. There are plans to extend the system to cover other Polynesian languages too. The developers, at Otago's Department of Computer Science and School of Maori Studies, are also looking at applications that tap into the lucrative market in computer-aided language-learning systems for English.

Why develop LT resources for a 'minority' language like Maori? The reasons are obvious from the point of view of the language itself. As LT becomes more prevalent, an LT infrastructure for a minority language could make all the difference to its continued existence. However, there are also compelling theoretical reasons. Maori, and its Polynesian cousins, are languages whose properties have intrigued theoretical linguists for a great many years. Development of a Maori system may generate insights that influence language technology for more commercially mainstream languages. To find out more, visit <http://tutoko.otago.ac.nz:8080/teKaaito>.

Who gets LT Update?

LT Update is a product of Macquarie University's unique teaching program in the human language technologies. This program, funded under the Federal Government's prestigious Science Lectureships Initiative, is the only teaching program in Australia that focuses on delivering a rich education in the twin areas of spoken language processing and natural language processing, widely viewed as critical technologies in the next few years. LT Update is provided as a service for alumni from this program, so it provides both a community for those with similar interests, and at the same time a very focussed channel to a group of people with particular skills. Thanks to CSIRO's generous support, subscriptions are currently free: visit <http://www.clt.mq.edu.au/LTUpdate> to register. You can also access this newsletter electronically via the site and you'll also find there web links to all the items mentioned in this issue as well as pointers to further resources.



LT update

what's happening in speech
and language technology in Australasia

CONTENTS

- Welcome to the third issue
- Tech Focus: Controlled Languages
- What is Intelligent Text Processing?
- Appen wins awards
- Language Technology goes Polynesian
- Upcoming Events
- Who gets LT Update?

Welcome to the third issue of LT Update!

LT Update is a free publication from the Centre for Language Technology, produced with the generous support of CSIRO. The Update is a twice-yearly hard and soft copy publication, backed up by timely email alerts, that aims to keep you abreast of developments in the speech and language technologies in Australia and New Zealand.

If you're not yet a subscriber, sign up at <http://www.clt.mq.edu.au/LTUpdate>. If you are a subscriber and you want to change your subscription details, visit the site and key in the six-character passcode printed on the top of your mailing label.

In past issues of LT Update, we have focussed on speech technologies. Speech processing, however, is just one side of Language Technology: there is also a significant commercial interest in intelligent technologies that work with text. So, in this issue, we look at what we call intelligent text processing: linguistically-aware software technologies that achieve results by manipulating, and in some sense 'understanding', text. If what's here piques your interest, you can find out more via the links for this issue at our website: visit <http://www.clt.mq.edu.au/LTUpdate>.

What's your view ?

If you have comments on LT Update, or ideas on things you'd like to see us cover, just mail ltupdate@ics.mq.edu.au.

Tech Focus: Controlled Languages

In each issue of LT Update, we bring you a brief primer on an important area of speech and language technologies. In this issue, Rolf Schwitter provides an introduction to controlled languages.

Full natural language is an indomitable beast when it comes to machine processing. Ambiguities and vagueness that humans handle with ease lurk in most sentences and make the automatic processing of natural language extremely difficult. The problem becomes particularly acute when texts need to be translated into other languages.

To overcome this problem and to reduce human translation costs, the Caterpillar Tractor Company (USA) introduced the first controlled natural language for machine translation in the 1960s. Today, Caterpillar uses the KANT machine translation system (<http://www.lti.cs.cmu.edu/Research/Kant/>) together with a modern version of a controlled natural language for the entire document production process.

In essence, a controlled language is a subset of a natural language that has been restricted with respect to its grammar and its lexicon. Grammatical restrictions result in less complex and less ambiguous texts. Lexical restrictions reduce the size of the vocabulary and the meaning of the lexical entries for a particular application domain. These restrictions improve the readability and clarity of the source text and produce better results with machine translations.

Controlled natural languages have also been designed for other purposes other than machine translation. Probably the best-known controlled language is AECMA Simplified English (<http://www.aecma.org/Publications.htm>). This controlled language was developed to facilitate the use of aircraft maintenance documentation by non-native speakers, and has been adopted as the de facto standard by the entire aerospace industry. Nowadays, aerospace manufacturers are required to write aircraft maintenance documentation using AECMA Simplified English.

In general, controlled natural languages are easier for humans to understand and easier for computers to process. However, these benefits

come at a cost, since controlled languages are difficult to learn and to remember. To improve the acceptability of controlled languages among technical authors, sophisticated language engineering tools such as grammar and terminology checkers have been developed. Grammar checkers such as Boeing's Simplified English Checker (<http://www.boeing.com/phantom/sechecker>) check input strings against the grammar rules and flag unapproved constructions. Terminology checkers help authors to add new vocabulary to a terminology database and to check the consistency of the terminology within documents. This is of particular benefit when a large number of authors contribute to a document.

At Macquarie University, research is under way that investigates how controlled natural language can be used to write specifications that bridge the gap between informal and formal descriptions. PENG (Processable ENGLISH) is a computer-processable controlled natural language that can be used to write precise and unambiguous specifications for knowledge representation (<http://www.ics.mq.edu.au/~rolfs/peng/>). The restrictions of this controlled language allow authors to express specifications in a familiar natural language notation and to combine this with the precision of a formal specification language. The author does not need to learn and to remember the restrictions of the controlled language. ECOLE, a sophisticated look-ahead editor, shows which structures are admissible. The resulting specifications look informal but have the same formal properties as the underlying formal language. Our experiments show that PENG is easy for non-specialists to write and understand, and easy to translate into a formal language. The underlying formal properties of PENG make it possible to use off-the-shelf theorem provers and model builders to check texts for inconsistencies. PENG can be adapted for other purposes that require precise input, e.g. for writing ontologies (PENG to RDF), for knowledge acquisition, or even for teaching students logic.

Find out more about controlled languages by following the links at the LT Update site.