



■ *India is a multilingual and multicultural country that came together less than a century ago. The populace spans wide extremes of wealth and education. The artificial intelligence community, which gained in strength in the 1980s, has had a major focus on research directed toward societal goals of bridging the linguistic and educational divide, and delivers the fruits of information technology to all people. In this article we look at a brief history followed by two examples of research aimed at crossing the language barriers.*

Deepak Khemani

A Perspective on AI Research in India

Artificial intelligence in India has been pursued by a passionate few over the last few decades. It has not been as widespread as in Europe and the USA. This could be due to two reasons. One is that research groups in general have been slow to gain in strength and have typically formed around a few diehard individuals scattered across the country. The priority in the first 50 years of independent India had been on undergraduate engineering education, and during this period students had inevitably gone westwards for doctoral studies, often staying back. The second was the propensity of the computing industry toward more lucrative assignments in the service sector. Both these factors are changing, not least because leading international software companies have set up research and development centers in the country.

Computer science education established itself in India in the early 1980s when the Indian Institutes of Technology (IITs) set up computer science departments and started offering undergraduate programs in the discipline. Research in artificial intelligence took off soon afterward when the government of India launched the Knowledge Based Computing Systems (KBCS) program in conjunction with the United Nations Development Program (Saint-Dizier 1991). A number of nodal centers were set up to focus on different areas of research including expert systems (IIT Madras), speech processing (Tata Institute of Fundamental Research), parallel processing (Indian Institute for Science), image processing (Indian Statistical Institute), and natural language processing (Center for Development of Advanced Computing).

Some of the early research in AI was motivated by societal needs. A prime example of this is the system Eklavya, a knowledge-based program designed to support a community health worker in dealing with symptoms of illness in

toddlers (Chandrasekhara, Shanthi, and Mahabala 1994). The program, named after a disadvantaged character in the epic Mahabharata, was meant to help create systematic case histories, provide basic treatment advice, or indicate the need for a referral, much like the call center software deployed in more modern times. Other examples are the language teaching system Vidya, the flight scheduling expert system Sarani (developed at CDAC, Mumbai), and a speech synthesis system developed for the railways by TIFR. Some of the people working in the KBCS projects found their way into the software industry and were instrumental in seeding in-house projects. For example Vivek Balaraman and his team at Tata Research Centre Pune developed a case-based reasoning kernel.

In the rest of the column we take a closer look at two strands of AI research in India that have thrived in the post-KBCS era.

Machine Translation

India is home to hundreds of different languages, with 22 being designated as official. Given that an average individual is familiar with only a few, machine translation (MT) and more recently cross language information retrieval has been a magnet for researchers. One group that has moved significantly beyond toy demonstrations is led by Rajeev Sangal and Vineet Chaitanya. The group, which had its genesis at IIT Kanpur in the 1980s, is currently active and growing at IIIT Hyderabad.

Fully aware that machine translation is a hard problem, the group embarked on translation with a basic system called Anusaaraka. The system exploited the fact that many Indian languages have well defined ways of depicting case markers by inflexions on words (Bharati, Chaitanya, and Sangal 1994). Coupled with the fact that these explicit markers, called *vibhakti*, can be mapped across languages, *Anusaaraka*, which means “the conformist,” is designed to do a translation in which the reader actively brings to the fore her or his world knowledge to quickly get a gist of the content. The emphasis is on comprehensibility and access to content as opposed to grammatical correctness. The system is not self-contained, which means that for tasks like literary translations the output will have to be postedited by a human for grammar and style.

The quality of a natural language processing system is directly dependent of the linguistic data that the system has access to. To this end the government of India is supporting the Linguistic Data Consortium for Indian Languages (LDC-IL) spearheaded by Sangal. The goal is to create a “repository of linguistic resources in all Indian languages in the form of text, speech and lexical corpora.”

The utility of such corpora is demonstrated by

the automatic translation system Sampark launched recently by the consortium (Anthes 2010). Sampark adopts a three-phased strategy for translation. In the analysis phase a series of modules (tokenizer, morphological analyzer, part of speech tagger, chunker, named entity recognizer, parser, and word sense disambiguator) generate an intermediate representation that is easy to transfer (syntax transfer, lexical transfer, and some transliteration). In the third phase target generation takes into account the agreement between phrases and insertion of the appropriate case markers. Sampark employs a mix of rule-based and statistical machine-learning approaches for different modules and exploits large amounts of linguistic data created by many teams of lexicographers working in eight Indian languages and English across the country. Currently translation between 18 language pairs is being developed.

Lexico-Semantic Relations

A key problem in natural language processing is word sense disambiguation. The Wordnet, a taxonomy of synsets (synonym sets) and relations between them, can be used for this task. The first one in English was created in Princeton. A Wordnet may be created manually by lexicographers by first principles. An alternative, as was first done for EuroWordnet, is creating synsets in a target language by mapping them from synsets of a source language. The first principles process is slow but produces synsets that have high fidelity to the concepts embodied by a society (in their language). The second is faster but may miss out on target language concepts. For example, the concept (represented by the word) *nephew* in English is further split into two concepts *bhatijaa* (brother's son) and *bhanjaa* (sister's son) in Hindi. In turn the concept *bhatijaa* is further split into *elder brother's son* and *younger brother's son* in Telugu. Indian languages contrariwise cannot describe the many forms of snow that Inuktitut would.

An approach is to borrow as much as one can from other Wordnets, and it works well for languages that are similar. A team led by Pushpak Bhattacharya at IIT Bombay started with a Marathi Wordnet created by expansion from the Hindi Wordnet that they had created by first principles (Bhattacharya 2010). Following this various universities across the country have come together to build cross-linked Wordnets in 16 of the 22 official languages. These are, Hindi, Marathi, Konkani, Sanskrit, Nepali, Kashmiri, Assamese, Tamil, Malayalam, Telugu, Kannad, Manipuri, Bodo, Bangla, Punjabi, and Gujarati, apart from Urdu. These languages come from diverse origins ranging from Indo-Aryan to Dravidian to Sino-Tibetan. Initially 32,000 synsets were assigned to six persons to

Worldwide AI

We invite AI researchers and professionals to submit columns on major international activities to share with readers of AI Magazine.

We hope that you enjoy and will contribute to "Worldwide AI!"

categorize them into four categories. Of these, 16,000 common synsets were identified and cross-linked in the Wordnets. A similar exercise to link these synsets to English has also been carried out.

From Text to Speech

The government of India is aggressively pursuing text to speech as a means to help those with visual impairments as well as those in rural areas with no access to the written word. To that end, a consortium to develop text-to-speech synthesis systems for Indian languages was formed in April, 2009 — Text to Speech Synthesis Systems for Indian Languages (TTS-IL). IIT Madras, led by Hema A. Murthy's group, was selected to develop speech synthesizers for six different Indian languages, Tamil, Hindi, Marathi, Telugu, Bengali and Malayalam. The synthesizers are integrated into screen readers (NVDA and ORCA) that can be used by the visually challenged. Other participating organizations are IIIT Hyderabad, IIT Kharagpur, CDAC Mumbai, and CDAC Trivandrum.

To enable disabled students to learn to use computers using the software, short term courses have been held at various institutes. In 2011, for example, workshops were held in Hindi at Saksham Delhi; Telugu at IIIT Hyderabad; Tamil at IIT Madras; Malayalam at CDAC Trivandrum; Bengali at NAB Kolkata; and Marathi at CDAC Mumbai. At these workshops, students were taught to how to use computers in their native languages, handle the keyboard, use a word processor, and access the internet, e-mail and spreadsheet; all using audio feedback provide by the text-to-speech systems (see for example Venugopal 2011).

A parallel effort is in the area of automatic speech recognition. Umesh Srinivasan's group at IIT Madras was chosen by the Department of Information Technology to lead a consortium that would develop systems to access agriculture prices via mobile telephones. Prices for various commodities are obtained daily from the Department of Agriculture and Cooperation's website. The information is then made available by telephone in local languages. The system performs limited vocabulary recognition and is available in six different languages — Assamese, Marathi, Hindi, Telugu, Tamil, and Bengali.

Concluding Remarks

Much of AI research in India has been driven by the need to bridge the language barriers in the country and also to enable disadvantaged sections of the society to reap the benefits of information technology. Following in the wake of an explosion in mobile phone penetration, there is considerable work being done in making useful information accessible to people in different languages and through alternate means, for example, accessing the web for the visually challenged. In this article we have looked at some language technologies being developed in India. There is more being done in AI. We hope to look at some of it in the future.

References

- Anthes, G. 2010. Automated Translation of Indian Languages. *Communications of the Association for Computing Machinery* (CACM) 53(1): 24–26.
- Bharati, A.; Chaitanya, V.; and Sangal, R. 1994. Paninian Framework and Its Application to Anusaraka. *Sadhana* 19(1): 113–127.
- Bhattacharya, P. 2010. IndoWordNet. In *Proceedings of the International Conference on Language Resources and Evaluation*. Paris: European Language Resources Association.
- Chandrasekhara, M. K.; Shanthi, B.; and Mahabala, H. N. 1994. Can Community Health Workers Screen Under 5 Year Children with Computer Program. *Indian Journal of Pediatrics* 61(5): 567–570.
- Saint-Dizier, P. 1991. The Knowledge-Based Computer System Development Program of India: A Review. *AI Magazine* 12(2): 33.
- Venugopal, V. 2011. A New World Opens Up. *The Hindu* 27 June. (www.hindu.com/2011/06/27/stories/2011062760780200.htm)

Deepak Khemani is a professor of computer science at IIT Mandi on leave from IIT Madras. His interests are in knowledge representation, case-based reasoning, planning, and qualitative reasoning. His long term-goal is to build articulate intelligent systems.