[From: *International Journal of Translation*, vol.13, no.1-2, Jan-Dec 2001]

CRL: A Center of Research Excellence in Machine Translation

SERGEI NIRENBURG

The Computing Research Laboratory (CRL) at New Mexico State University (NMSU) is a non-profit, self-supporting research enterprise committed both to basic research and to development of software systems.

CRL's research efforts concentrate on practically all approaches to multilingual processing of natural language texts. Fields of study contributing to this endeavor, and thus among the core areas of research at CRL, are artificial intelligence, computational linguistics and human-computer interaction. This document provides brief glimpse into CRL's current and recent activities. More information about CRL can be found at its website: http://crl.nmsu.edu.

RESEARCH SPANNING THE DISCIPLINES

CRL pursues *research* in computational morphology, computational syntax and ontological semantics, including development of large computational ontologies and interlingual text meaning representation languages. The lab develops *systems* for morphological, syntactic and semantic analysis of texts, including treatment of reference and non-literal language (e.g. metonymies, metaphors, etc.); text planning and text generation; storage, search and control architectures; and user and knowledge acquisition interfaces.

CRL's R&D teams develop and deploy proof-of concept and proto-type systems for machine translation, information retrieval (IR), information extraction (IE), summarization, question answering and other applications. IR and IE systems at CRL can be configured for single-language and cross-lingual operation.

CRL has developed machine translation systems using practically every extant technique (often, the systems feature a combination of such techniques). CRL's experience in machine translation includes the ULTRA and Mikrokosmos knowledge-based machine translation projects; the Pangloss project for knowledge-based, machine-assisted translation; development of a number of transfer-based MT systems; the Temple Translator's Workstation project; Corelli human-assisted machine translation project, the Cibola translation support system and the Oleada environment for language learning and translation assistance.

Cross-lingual information retrieval, text summarization, information extraction and question answering are studied at CRL using a variety of techniques and are configured in a variety of integrated systems, for instance, in a prototype system supporting a team of human and software agents in preparing reports about emerging crises. CRL is a regular participant in the MUC, TREC, and TIPSTER evaluations. CRL is the only TREC participant to actually test systems in which English queries are used to retrieve documents in multiple languages.

Data Resources at CRL

The work on both the technologies and applications is supported by a massive effort to acquire computational resources, such as dictionaries, grammars, corpora and other data. CRL has developed a substantial archive of data resources to support its research efforts. This archive is continually growing with each new project. CRL has acquired resources in over 20 languages.

A Brief Survey of Current Research Projects

Recent basic research and development activities at CRL have concentrated on the development of methods for producing robust large-scale natural language processing system prototypes. In what follows, we give the reader a brief tour of current R&D at CRL.

CAMBIO

CRL's CAMBIO project is devoted to building an environment for facilitating augmentation of lexical knowledge by people who have not - unlike the dedicated developers of such systems - been specially trained for the task, for example, intelligence analysts.

The work on enhancing the lexicons and the ontology is ongoing over the life of any knowledge-based system. In particular, one must expect the appearance in input texts of lexical units whose meanings have not yet been adequately captured by the system. This can happen when the lexical unit had not been attested at all in input texts or when it had not been attested in the particular sense that is relevant to the text in question. For example, the English spring could be known to the system's lexicon in the sense of a mechanical device but not in the senses of the season or source of water.

The CAMBIO project is intended to investigate multiple aspects of knowledge-based support for intelligence operations. In Phase II, CRL will investigate the potential of knowledge-based document retrieval in a multi-lingual document collection. This will be compared against a baseline using a bilingual lexicon technique (also developed at CRL). The knowledge base used will be the Mikrokosmos ontology. The input to the augmentation process in the proposed work will be a list of lexical units that have to be included in the lexicon (often causing the need for modifications in the ontology). We propose to build a lexicon acquisition interface and an ontology acquisition interface to support this task. The most important feature of these *integrated* interfaces will be their reliance on the methodology of knowledge elicitation developed in the Expedition project at CRL. Expedition leads the untrained user through a series of acquisition steps by the system, with help, material tutorials and explanations available at every step.

CREST

The effort in the CREST (Crosslingual Retrieval, Extraction, Summarization and Translation) project is focused on development of working systems for a number of natural language applications including machine translation, crosslingual information retrieval, crosslingual and cross-document information extraction, summarization and task-oriented combinations of the above.

The main thrust of the research is toward enhancing basic capabilities in text processing, with the view of attaining a new level of quality for NLP applications. The main thrust of the development is toward judicious integration of diverse methods and resources in application-oriented systems.

Building on the NLP resources, formalisms, techniques, system modules, architectures and tools accumulated over the years at CRL, the project targets fundamental advances in semantic analysis of text. In particular, it concentrates on full-text word sense disambiguation, encoding of sentence and text meaning, and treatment of non-literal language (metaphors, metonymies and other tropes). Additionally, advances will be sought in coreference text analysis and text generation.

The end goal is to develop a *meaning-based* environment to support a team of human and software agents in performing cross-lingual text processing tasks.

Expedition

This project is devoted to developing a computational environment for enabling quick ramp-up of machine translation systems for "low density" languages, that is, languages for which large-scale machine translation systems have not been developed. This is a large effort that centrally involves developing a knowledge elicitation system to allow a speaker of any of about 60 designated languages to provide the system with knowledge about its morphological and syntactic properties as well as create a bilingual dictionary between that language and English.

A special direction of work is researching new algorithms for gradual and partial automation of such knowledge acquisition tasks. In addition to the above, the project involves testing the nascent elicitation system by actually developing machine translation system prototypes for three concrete languages (one for each of the three years of the project). More information about Expedition can be found at http://crl.nmsu.edu/Expedition

GraphLing

This work involves developing probabilistic classifiers for two challenging and diverse natural language processing tasks using a common set of techniques. One classifier will be capable of disambiguating a large vocabulary of words with respect to a full set of sense distinctions from a published source, such as Longman's online dictionary. The second will perform a discourse processing task that involves segmentation, reference resolution and belief: segmenting a text into blocks that express the beliefs and opinions of a single agent, and identifying noun phrases that refer to that agent. Both systems will be fully automatic.

URSA

URSA is CRL's TIPSTER Phase III research and development effort to make text processing and information retrieval transparent to languages.

URSA combines Unicode display technology developed at CRL with translingual information retrieval, multilingual collection visualization and document management, with special emphasis on design principles that have been validated by examining the analyst in real-world scenarios.

URSA combines the latest advances in information retrieval with the coherent Unicode text model to make language-transparent IR a reality. Ongoing development is focusing on integrating Unicode detection technology with the Tipster architecture. The model we are developing utilizes annotations on the documents to describe the text for indexing. External document annotators can produce segmentations for Oriental languages, or stemmed word markup for Western languages, which are then interpreted by the URSA engine and indexed for later retrieval. The URSA engine will be fully conversant in Tipster detection needs, including complex query expressions and natural language queries. More information about URSA can be found at http://crl .nmsu.edu/Researoh/Proiects/tipster/ursa.

RECENTLY COMPLETED RESEARCH PROJECTS

Artwork III

Artwork addressed the machine translation of spoken dialogue. The focus was investigating approaches to providing robustness by exploiting models of the task domain and of conversational interaction to generate relevant expectations against which the input can be interpreted.

Spoken dialogue, as opposed to edited text, is characterized by short unedited utterances that are less likely to conform to a standard language than edited text. A dialogue translation system will be much more dependent on robustness techniques. The goal of Artwork was to recover enough of the speakers' intended meanings from their utterances to be able to generate satisfactory English translation.

ATT-Meta

This artificial intelligence project investigated part of the problem of discerning the coherence relationships within natural language discourse. The relationships of interest were between discourse components that describe or presuppose mental states (propositional attitudes). The mental states could be those of the speaker, listener, or any person mentioned by the discourse and could be states of belief, intention, hope, desire, etc. The focus was on descriptions that relied implicitly or explicitly on common-sense models of the mind that people often deploy, such as the prevalent model of a mind and its beliefs as, respectively, a physical container and some physical objects within it. The project included the development of a prototype artificial intelligence discourse-interpretation system. This system included a parser, a metaphor recognizer, and most important, a default-inference module. An episodic logic including defaults was developed for use by the inference module. The project also involved major data collection activities: collection of information about mental-state language texts and of definitions and examples in machine-readable dictionaries. Annotated compendia of examples from these texts were produced.

Cervantes

This research continued the DARPA-sponsored Tipster project on information retrieval. It was a joint effort among many sites to develop a working system that integrates information retrieval and information extraction. At the core of the effort was a joint government/contractor committee that is specifying an architecture for this kind of system. Two CRL principal investigators are members of this committee.

In addition to further development of the Diderot information extraction system developed in phase one of this project, CRL provided a variety of specialized software subsystems to support the architecture development.

Cervantes has major implications for the construction of information analysis systems that process large volumes of textual data. In particular, it has long-term potential for the creation of large-scale text databases; new applications where considerations of speed of data capture and analysis outweigh the need for 100 percent accuracy; retrieval against multilingual text data bases; and user interface support targeted at analysis and related problems.

Corelli

The goal of the Corelli project was to develop a framework, architecture and tools, for rapid deployment of multilingual machine-translation systems, with an emphasis on machine translation for assimilation purposes and on languages for which electronic or human resources are scarce or difficult to obtain. Building on previous achievements from the Temple project, the Corelli project concentrated on several areas: a general architecture for text engineering, finite-state transducer technology, multilingual lexical tools and application to new languages. While language components developed in the Temple project (Arabic, Japanese, Russian and Spanish to English) were ported to the new architecture, new Serbo-Croatian and Korean components are being developed. CRL Project Manager Remi Zajac led the Corelli project. More information on this effort can be found at http://crl.nmsu.edu/Research/Projects/corelli.

Keizai

In this project CRL integrated retrieval and summarization technologies in a demonstration system called Keizai (Japanese for “economy”). Keizai accepts English queries; retrieves documents in Japanese, Korean and English, and other languages; produces summaries of documents and document sets in English; and displays the documents and summaries in a variety of ways of interest to the economic analyst. The user interacts with the system via a WWW-browser with most of the significant components of the prototype system being hosted through a Web server. Keizai combines the cross-language retrieval capabilities of the URSA project with the multilingual summarization abilities of MINDS, as well as introduces new capabilities for displaying the relationships between summarized documents and econometric data. The retrieval technologies of URSA are based on indexing and presenting documents in Unicode, a method for encoding most of the world's languages. In URSA, queries are translated at query time into the language of the target documents. Research has investigated the performance of bilingual dictionary-based query translation and the impact of using parallel text resources to enhance the performance of dictionaries by, for example, translating domain-specific terms that don't occur in dictionaries. The findings over the past several years suggest that query translation can be highly effective, but that parallel text resources must be deployed sparingly because of the high degree of noise and ambiguity inherent in parallel texts.

Keizai leverages CRL's multilingual resources and onomastica (proper-name dictionaries) that have been developed in the context of MINDS and other projects to assist in translation of queries and summaries.

Mikrokosmos

Mikrokosmos was devoted to a study of computational semantics of natural language, the results of which were tested in a knowledge-based, interlingual MT system from Spanish, Chinese and Japanese into English. Methodologically, Mikrokosmos was based on the recognition that computational treatment of text requires the study of a wide variety of language, language use, and world phenomena and that a single all-encompassing theory of computational linguistics is not feasible, at least in the near future. Natural language application grounded on a sound theoretical basis is to devise a computational architecture that allows for the integration of partial treatments, i.e., microtheories, of a variety of language phenomena. The integration is carried out through a uniform knowledge representation formalism and a flexible control architecture in the application system. Mikrokosmos centrally addresses microtheories of lexical meaning, semantic dependency, aspect, time, reference and style. In its framework semantic analyzers for the source languages and a sentence planner for English have been developed. CRL Director Sergei Nirenburg led Mikrokosmos. More information about Mikrokosmos can be found at http://crl.nmsu.edu/Research/Projects/mikro.

MINDS

The MINDS (multilingual interactive document summarization) project involved the summarization, extraction and translation for Japanese, Spanish and Russian. The MINDS project focused on the creation of a multilingual summarization tool designed to provide quick and interactive document filtering, even in the absence of certain lexical or other resources for a language. MINDS was a DARPA-sponsored Tipster Phase III project.

MINDS technology, for English, has now been evaluated formally as part of the Tipster program. Humans engaged in a classification task that processed the summaries produced. A significant reduction in the time to read documents was observed, with only a small reduction in the accuracy of the judgments made. MINDS was also engaged in developing goal-based summarization that extracts information on specific event types from texts. This information is then used to generate summaries that link together all documents related to an event. The final cross-document summary will be in English with links to summaries in English produced by machine translation and to the full texts in the original four languages. More information about MINDS can be found at http://crl.nmsu.edu/Research/Projects/minds

Oleada

This project was an extension to the technology, resource base and graphical user interface design developed in the Cibola project. Intended both for language instructors and language students, Oleada enhanced instructors' abilities to create and modify language instruction modules and improved learners' proficiency through self-study and evaluation. This project uses both information extraction and document detection technology to provide language analysis tools. More information about these projects can be found at http://crl.nmsu.edu/Research/Projects/oleada

Pangea

Begun in September of 1997, the Pangea project was a three-year effort to create a standards-based infrastructure for multilingual computing in general as well as natural language processing. This effort produces the Multilingual Unicode Text Toolkit, or MUTT, for application developers. CRL has provided the government with a toolkit that is widely used, supporting hundreds of users in a variety of languages. Some of its strengths include a wide variety of easily extensible input methods, full Unicode text editing in the Motif/X11 -based environment and initial Java support. MUTT currently supports conversion between Unicode and more than 100 character sets and transliterations.

Pangloss

This project encompassed the investigation and development of a new generation of machine translation systems within a knowledge-based, interlingual approach and combined this approach with others in a single system. The project involved CRL, University of Southern California and Carnegie Mellon University. Work at CRL included building high-quality syntactic, semantic and pragmatic analyzers for Spanish and acquiring knowledge needed by the system, both automatically and manually. Knowledge acquisition was used to build lexicons (English and Spanish), acquire a domain model (ontology) and develop an interlingual concept lexicon. The Spanish analysis system (the Panglyzer) is a modular system, with results at each

Stage feeding into the subsequent module. The resultant readings are ranked and the output is passed to the generation system.

Shiraz

The goal of this project was the creation of an extensible research prototype of a Persian to English machine translation system that can be used to explore the requirements of future Persian machine translation development. The approach enables users to pre-process Persian texts using Persian processing tools that incorporate knowledge of the Persian language, thus relieving users from low-level pre-processing and correcting tasks. Users can consult, modify and enrich the machine translation system dictionaries through specially designed user-interfaces for processing new textual material. Finally, the integrated system enables users to access and manage documents to be translated and their translations, including the export of translated text to word-processing software. More information about Shiraz can be found at http://crl.nmsu.edu/Research/Projects/shiraz.

STRATEGIES FOR NEW TECHNOLOGY

Although CRL is an academic lab, its central output includes a variety of working and deployed prototype systems for the applications studied - well beyond the publications usually expected from an institution of this kind. Research prototypes are often developed with substantial effort to address the kinds of tasks that a system would be used for in real-world settings, providing a flexible environment that allows for easy integration, adaptation to specific needs, and the opportunity to upgrade to the newest most advanced technologies.

CRL'S RESEARCH AND DEVELOPMENT STRATEGIES

- Concentration on large-scale, *multilingual* text and speech processing.
- All research results are delivered in *working systems.*
- R&D methods and techniques are chosen from the point of

view of *benefits for the task* at hand; no "have method, will travel" attitudes.
- Emphasis on utility of research results leads to incorporation of human agents into many system environments.
- Attention to methodology *of massive* knowledge acquisition.
- Strong preference for close contact with potential and actual users at all stages of system design and development.
- Insistence on *mixing* short- and long-term R&D goals.
- Preference for *combining* R&D methods and techniques in a single system; this includes also reuse of available modules and use of commercial off-the-shelf software.

CRL RESEARCH FACULTY AND STAFF

CRL is a team of more than 70 scientists, technicians, graduate research assistants, expert consultants and support staff. Currently the lab employs nearly 20 Ph.D.-level researchers. Their work addresses complex problems from an interdisciplinary perspective with computer scientists, engineers, linguists, mathematicians and psychologists. In addition to their research expertise, CRL staff bring to the laboratory native fluency in seventeen major world languages including-besides English-Arabic, (Mandarin) Chinese, French, German, Farsi, Hebrew, Hindi, Japanese, Korean, Russian, Serbo-Croatian, Spanish, Thai, Turkish, Ukrainian and Urdu.

Following are brief biographies of CRL's leadership staff and key researchers. More information about these and other CRL researchers can be accessed via the CRL Staff web page at http://crl.nmsu.edu/Staff/crlstaff.htm.

Recognition through Publishing

CRL's work is recognized around the world through the prolific work of its researchers. On average, CRL staff and adjunct researchers produce more than 50 conference papers, presentations and journal articles each year. Additionally, CRL publishes *Memoranda in Computer and Cognitive Science.* A bibliography of CRL's publishing efforts can be found at http://crl.nmsu.edu/Publications/publicat.htm.

Conference Participation around the Globe

The research teams at CRL are actively pursuing research in most of the current paradigms of natural language processing, artificial intelligence and graphical user interface design. Its interdisciplinary team of computer scientists and engineers, linguists, mathematicians, psychologists and electrical engineers are active participants in national and international conferences and committees. CRL staff members share their expertise worldwide by taking leadership roles in planning, sponsoring and presenting new research findings at conferences on a regular basis. Information about CRL staff conference participation can be found at http://crl.nmsu.edu/Events/external.htm.

During the summer of 1999, CRL conducted the Summer School in Language Engineering at the NMSU campus. Some *25* participants from origins worldwide attended two-weeks of intensive instruction and hands-on practical application in such areas as machine translation, information retrieval and extraction and text summarization. Courses were taught by CRL researchers and visiting scholars. Courses included "Ecological" Issues in Language Engineering, Lexicon Acquisition for NLPI: Morphology and Syntax, Lexicon Acquisition for NLP II: Ontological Semantics, Approaches to Computational Morphology, Knowledge Elicitation from Informants, and A Survey of Language Engineering Applications. Some of the materials presented at the summer school can be found at http://crl.nmsu.edu/resources/materials.htm.

In the summer of 2000, CRL, along with Bilkent University and Tbilisi State University, cosponsored the NATO Advanced Study Institute on Language Engineering for Lesser-Studied Languages (ASI). Held on the campus of Bilkent University in Ankara, Turkey, the program comprised of state-of-the-art courses on various aspects of language engineering.