Machine Translation, Digital Libraries, and the Computing Research Laboratory

Stephen Helmreich
Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003 USA
shelmrei@crl.nmsu.edu
http://crl.nmsu.edu
(505) 646-2141

Machine Translation (MT), along with Natural Language Processing (NLP), is one of the oldest problem areas in that branch of computer science known as Artificial Intelligence. Over its history, it has fostered a number of technologies, resources, and tools, and, now (through the adjective "multi-lingual") finds itself incorporated into other language-related tasks such as Information Retrieval (IR), Text Summarization, and Information Extraction (IE), known collectively as Human Language Technologies (HLT).

Current research now focusses on combining these technologies, along with similar technologies in other information media (speech recognition, vision, video, structured and unstructured data, knowledge-representation and reasoning) into integrated information-technology (IT) programs, such as question-answering systems, personal profilers, or intelligent tutoring systems.

In this presentation I focus on those tools, resources, and areas of research in this very broad area that might be of value to developing and enhancing the value of Digital Libraries, and particularly those that might be of value in Indo-US collaboration. I will exemplify these, where possible, with projects and resources developed at the Computing Research Laboratory (CRL), though there are numerous institutions engaged in this research. (In the paragraphs below, I have put in bold-face those areas of great interest, and in italics, those of possible interest.)

Machine Translation (Purposes): *(1) Dissemination – high quality, sublanguages, controlled languages; (2) Assimilation – broad coverage, lower quality;* (3) Communication – rapidity, flexibility.

Machine Translation (Types): (1) Direct – string-for-string, example-based; (2) Transfer – structure-for-structure (Systran); **(3) Interlingual – analysis to and generation from a meaning representation;** (4) Statistical – the most probable translation given a bilingual corpus.

Machine Translation Technologies and Resources: (1) character encoding and representation, text editing (Unicode, Multilingual Unicode Text Tool --MUTT); *(2) text segmenting (OCR, sandhi?);* **(3) Morphological analysis;** (4) Lexical annotation (part of speech, **proper name identification**); (5) syntactic analyzers (grammars, parsers); **(6) bilingual/multilingual dictionaries; (7) ontologies (OntoSem, WordNet, Cyc);** (8) Generation systems.

Human Language Technologies: (1) Information Retrieval (I won't even go there!); (2) Information Extraction (Message Understanding Conferences – MUC); (3) Text Summarization (Document Understanding Conferences – DUC); (4) Word sense disambiguation (SensEval); (5) Cross Document Named Entity Identification.

Enhanced HLTs: All of the above with multilingual, multi-modal capacity.

High-Level Information Technologies: (1) Personal Profiler (IR, MT, Summarization) (2) Quick Ramp-up MT (Expedition – HCI, MT); (3) Personal Question-Answering Systems (Advanced Question and Answering for Intelligence – AQUAINT; Meaning-Oriented Question Answering MOQA – HCI, IR, IE, MT, Text generation).

In short, there are a wealth of resources, tools, and research areas in Machine Translation and related areas. There is no doubt in my mind that many of these resources and tools would prove useful as those involved in Digital Libraries face the difficult tasks of acquisition, maintenance, enhancement, and dissemination of digital information. In addition, looking at areas of research from both a digital libraries viewpoint and a language technology viewpoint should provide a fruitful method of cooperation and a viable source of funding.