AAMT

Asia-Pacific
Association for
Machine
Translation

Journal

12-16 SEPTEMBER 2005

MT SUMMIT X

PHUKET ✧ THAILAND

Special Issue
September 2005

CONTENTS

For Success of MT Summit X

MT Summit began at Hakone, Japan in 1987, which has been followed by eight successful conferences (Munchen, Washington, Kobe, Luxemburg, San Diego, Singapore, Santiago de Compostela, and New Orleans). This is,therefore, the 10th Summit and the fourth in Asia.

All of the preceding nine conferences were very successful and left distinctive impressions in the minds of the participants. I remember vividly the Summit at Santiago de Compostela immediately after 9/11, 2001, where we had many participants from the US who managed to come to the conference despite extreme difficulties in traveling. I also remember enjoyable conversations with Professor Antonio Zampolli (Pisa, Italy) at Singapore as well as Santiago, who passed away in August 2003, a month before the last Summit in New Orleans.

MT Summit X has also experienced sadness and difficulties. The conference venue, Phuket Thailand, suffered severe damage by the tsunami on December 26, 2004. About 2,000 were reported missing, dead and injured in Phuket alone. Despite of this monstrous human tragedy, we decided to hold the conference here at Phuket, because we have received massive support and encouragement from members of IAMT (International Association for Machine Translation) and because we believe the summit here would boost the moral of the local people and thus contribute to local recovery effort.

The MT summit has always attracted participants with very diverse backgrounds, from MT users to managers to engineers and researchers. It has been a unique conference where people who are interested in MT from whatever perspectives can come to exchange their views and ideas with others. Even in ordinary circumstances, to organize MT summit with such diverse participants is not easy. If MT Summit X is a success and I believe it certainly will be, it is due to devoted efforts by people who have strong wills to conquer the difficulties caused by the tragedy.

My special thanks go to the members of the Program Committee chaired by Dr. H. Isahara with Prof. B. Maegaard and Dr. L. Gerber as co-chairs, the members of the local organizing committee chaired by Dr. V. Sornlertlamvanich, the Co-Chairs for Exhbition (Mr. M. Ohori, K. Matsui, T. Murata), the Co-Chairs for Workshops (Prof. S. Kurohashi, Dr. E. Sumita), and the members of the Steering Committee. I also would like to express my gratitude to Prof. M. Nagao who initiated the first MT summit at Hakone 18 years ago and who kindly agreed to be the honorary chair of this conference.

Last but not least, I would like to thank Dr. M. Nakase and Ms. K. Takada of AAMT. Without their devotion, the life of the conference chair would be intolerable.

Junichi Tsujii
President of AAMT and IAMT
Conference Chair of MT Summit X
University of Tokyo, Japan
University of Manchester, UK

The History of the MT Summit

	Name	Organizer	Conference Site	Period
1st	MT Summit I	AAMT	Hakone, Japan	Sept. 17-19, 1987
2nd	MT Summit II	EAMT	Munich, Germany	Aug.16-18, 1989
3rd	MT Summit III	AMTA	Washington, USA	July 1-4, 1991
4th	MT Summit IV	AAMT	Kobe, Japan	July 19-22, 1993
5th	MT Summit V	EAMT	Hemicycle, Luxembourg	July 10-13, 1995
6th	MT Summit VI	AMTA	San Diego, USA	Oct. 29 - Nov.1, 1997
7th	MT Summit VII	AAMT	Singapore	Sept. 13-17, 1999
8th	MT Summit VIII	EAMT	Santiago de Compostela, Spain	Sept. 18-22, 2001
9th	MT Summit IX	AMTA	New Orleans, USA	Sept. 23-27, 2003
10th	MT Summit X	AAMT	Phuket, Thailand	Sept. 12-16, 2005

Venue of MT Summit X

Hilton Phuket Arcadia Resort & Spa

333 Patak Road, Karon Beach, Phuket, Thailand

TEL: +66-76-396-433 FAX: +66-76-396-136

URL: http://www.phuketarcadia.hilton.com

Regional Associations

Asia-Pacific Association for Machine Translation (AAMT)

http://aamt.info

Association for Machine Translation in the Americas (AMTA)

http://www.amtaweb.org

European Association for Machine Translation (EAMT)

http://www.eamt.org

【AAMT Board Members】

●President	Jun-ichi Tsujii (Professor, University of Tokyo)
●Vice President	Taizo Kotani (President, Inter Group Corporation)
●Director	Makoto Nagao (President, National Institute of Information and Communications Technology)
●Director	Hozumi Tanaka (Professor, Chukyo University)
●Director	Shun Ishizaki (Professor, Keio University)
●Director	Shoichi Yokoyama (Professor, Yamagata University)
●Director	Hitoshi Iida (Professor, Tokyo University of Technology)
●Director	Yoshiyuki Sakamoto (Professor, Tsukuba Women's University)
●Director	Hitoshi Isahara (Group Leader, Computational Linguistics Group, NICT)
●Director	Key-Sun Choi (Professor, Korea Advanced Institute of Science and Technology)
●Director	Virach Sornlertlamvanich (Director, Thai Computational Linguistics Laboratory)
●Director	Shigeji Kajikawa (President & CEO, Toshiba Solutions Corporation)
●Director	Chiaki Itoh (Member of the Board/Corporate Executive V.P., Fujitsu Ltd.)
●Director	Toru Chiba (Corporate Director and Group General Manager of Corporate Research and Development Group, Sharp Corporation)
●Director	Yosuke Takashima (General Manager, System Platform Software Development, Div., Solution Development Laboratories, NEC Corporation)
●Director	Manabu Shinomoto (Corporate Officer, Hitachi, Ltd.)
●Director	Harushige Sugimoto (Senior Vice President, Oki Electric Industry Co., Ltd.)
●Director	Masanori Fukiwake (President, Japan Electronics and Information Technology Industries Association)
●Auditor	Kazuaki Ogasawara (Executive Vice President, JEITA)
●Auditor	Mihoko Katsuta (CEO, Toin Corporation)

Asian Machine Translation Software Companies

AAMT Internet Working Group

★ We have listed the names of companies who sell MT software in Japan or MT software which translates from/to Asian Languages.
★ The listed sites are not related to AAMT and AAMT is not responsible for their content.
★ The list is presented in no particular order.

August, 2005

Sharp (http://www.sharp.co.jp/products/honyaku/)
Toshiba Solutions (http://hon-yaku.toshiba-sol.co.jp/)
TOSHIBA (http://www.toshiba.co.jp/index.htm)
NEC (http://www.sw.nec.co.jp/soft/crossroad-enterprise/)
Oki Electric Industry (http://www.yakushite.net/index.ja.html)
Fujitsu (http://software.fujitsu.com/jp/honyaku/)
Brother (http://www.brother.co.jp/jp/honyaku/honyaku.html)
JST (http://pr.jst.go.jp/pub/pubindex.html)
LogoVista (http://www.logovista.co.jp/)
Cross Language (http://www.crosslanguage.co.jp/)
PATOLIS (http://www.patolis.co.jp/)
MARUZEN (http://japanese.chosun.com/site/data/html_dir/2005/01/07/20050107000036.html)
Sakura (http://homepage1.nifty.com/tr/mouse/)
Impulse Japan (http://www.impulse-jp.net)
Fujitsu Learning Media (http://www.flm.fujitsu.com/)
A.I. Soft (http://ai2you.com/goma/)
Kodensha (http://www.kodensha.jp/)
IBM (http://www.ibm.co.jp/software/internet/king/)
IPM (http://www.ipm-c.co.jp/)
Omron Software (http://www.omronsoft.co.jp/SP)
Souiku (http://www.soiku.co.jp/products/h_hyi/index.html)
ASCII Solutions (http://www.asciisolutions.com/products/index.html)
Software Gijutsu (http://www.sofugi.co.jp/etranj/etjtop.html)
Logical tech (http://www.logicaltech.co.jp/LTCatsEye.htm)
Justsystem (http://www.ichitaro.com/option/transmas2/)
Yamano (http://www.tcct.zaq.ne.jp/yamano/index.html)
Word Bank (http://www.ashiya.ne.jp/rosetta.html)
Trilingual Peksong (http://www1.odn.ne.jp/~caa33950/myhome/)
TechnoWare (http://www.bekkoame.ne.jp/~twc/index.html)
Source Next (http://www.sourcenext.com/)
Unikotech (http://www.unikotech.com/)
Seagrand (http://www.seagrand.co.jp/products/eprinter/index.shtml)
NanaTech (http://www.nanatech.co.jp/nana1.html)
Device Net (http://www.devicenet.co.jp/pro/tabiec.html)
Media Drive (http://pac.mediadrive.jp/products/index.html)
Holon (http://www.holonsoft.co.jp/products/study/honyaku/index.html)
Accela Technology (http://www.accelatech.com/products/BL/index.html)

APRO Technology (http://www.aprotechnology.com/jp/products/ATransKJ/atranskj.html)
TechnoCraft (http://www.a2001.com/shop/roboword/new/index_dic.html)
i-4 (http://www.ifour.co.jp/press/n2005q1/20050304.html)
Canon (http://cweb.canon.jp/color-ir/index.html)
H. Takahashi (http://www.vector.co.jp/soft/win95/edu/se364617.html)
WAC.com (http://www.wac-jp.com/products/hangryu/)
IDENT (http://www.e-ident.net/htm_show.html)
Changshin Soft (http://cssoft.co.kr/jp/)
Haan Soft (http://www.haansoft.com/)
Worldman Corporation (http://www.worldman.com/)
Bencom Inc. (http://www.smartran.co.kr)
Dream C&C (http://www.dreamsell.co.kr/)
ClickQ Co.Ltd. (http://www.clickq.com/)
LNI Soft (http://www.nexosoft.co.kr/soft/product.asp?cate=tf)
sysmeta (http://www.sysmeta.com/)
Namsanjae Information Center (http://www.dprknta.com/business/trd/nanzankyu.html)
Kingsoft (http://www.iciba.net/)
Create Dalian (http://www.create-dl.com/chanpin.htm)
Computer and Microelectronics Industrial Deveropment (http://www.ciita.org.cn/)
Sunshine Technology Company Ltd. (http://www.sunshine-group.com/index.htm)
Transtar (http://www.transtar.com.cn/)
Microelectronics and Computer Development Center (http://www.chinatranslate.net/it/it300_19.htm)
Huajian (http://www.hjtek.com/)
Inventec (Shanghai) Co., Ltd. (http://www.dreye.com.cn/)
Hostran & Microc Software, Inc (http://www.hostran.com.tw/)
Otek International Inc. (http://www.otek.com.tw/)
National Center for Technological Progress (http://www3.thanhnien.com.vn/CNTT/KHtintuc-sukien/2005/6/6/111934.tno)
Kebutuhan Sistem (http://www.cdpenerjemah.cjb.net/)
Axel Blume (http://www.ablume.com/)
Agent Dict (http://www.agentdict.net/)
Centre for Development of Advanced Computing (http://www.cdac.in/html/aai/mantra.asp)
Padideh Co. (http://www.padideh.org/Englishindex.htm)
Mabnasoft (http://mabnasoft.com/english/parstrans/index.htm)
Arab.Net Technology Ltd. (http://www.arab.net/)
Babylon (http://www.babylon.com/)
Larry Smith (http://members.tripod.com/~Targumatik/)
ITC Inc. (http://www.itc.com.tr/engl/cev.html)
Bilsag Ltd. (http://www.bilsag.com.tr/)
Sakhr Software Co. (http://www.sakhr.com/)
Trident Software, Ltd (http://www.trident.com.ua/index.html)
Project MT Ltd. (http://shop.e-promt.ru/)
Cimos (http://www.cimos.com)
SYSTRAN (http://www.systransoft.com/index.html)
ATA Software Technology (http://www.atasoft.com/)
ArabNet Technology (http://www.gy.com/www/ww1/ww2/atabuot.htm)
Transparent Language (http://www.transparent.com)

World Language Resources (http://www.worldlanguage.com)
Civasoft Corporation (http://www.civasoft.com/products.htm)
Pacific Software Publishing, Inc. (http://www.pspinc.com/htm/jpn/jlsp-pro.htm)
ComCul International (http://www.comcul.com/denchan/index-j.html)
VirtualWare Technologies (http://www.allvirtualware.com)

History of Phuket

Phuket Island has a long recorderd history, and remains dating back to A.D. 1025 indicate that the island's present day name derives in meaning from the Tamil manikram, or crystal mountain.

For most of history, however, it was known as Junk Ceylon, which, with variations, is the name found on old maps. The name is thought to have its roots in Ptolemy's Geographia, written by the Alexandrian geographer in the Third Century A.D. He mentioned that in making a trip from Souwannapum to the Malay Peninsula it was neccesary to pass the cape of Jang Si Lang.

Phuket was a way station on the route between India and China where seafarers stopped to shelter. The island appears to have been part of the Shivite empire (called in Thai the Tam Porn Ling) that established itself on the Malay Peninsula during the first Millenium A.D. Later, as Muang Takua-Talang, it was part of the Srivichai and Siri Tahm empires. Governed as the eleventh in a constellation of twelve cities, Phuket's emblem, by which it was known to others in those largely pre-literate times, was the dog.

During the Sukothai Period Phuket was associated with Takua Pa in what is now Phang-nga Province, another area with vast tin reserves. The Dutch established a trading post during the Ayuthaya Period in the 16th Cent. The island's northern and central regions then were governed by the Thais, and the southern and western parts were given over to the tin trade, a concession in the hands of foreigners.

After Ayuthaya was sacked by the Burmese in 1767 there was a short interregnum in Thailand, ended by King Taksin, who drove out the Burmese and re-unified the country. The Burmese, however, were anxious to return to the offensive. They outfitted a fleet to raid the southern provinces, and carry off the populations to slavery in Burma.

This led to Phuket's most memorable hitoric event. A passing sea captain, Francis Light, sent word that the Burmese were en route to attack. Forces in Phuket were assembled led by the two heroines, Kunying Jan, wife of Phuket's recently deceased governer, and her sister Mook, After a month's siege the Burmese were forced to depart on 13 March, 1785. Kunying Jan and her sister were credited with the successful defense.

In recognition King Rama I bestowed upon Kunying Jan the honorific Thao Thep Kasatri, a title of nobility usually reserved for royalty, by which she is known today. Her sister became Thao Sri Sunthon.

During the Nineteenth Century Chinese immigrants arrived in such numbers to work for the tin mines that the ethnic character of the island's interior became predominantly Chinese, while the coastal settlements remained populated chiefly by Muslim fishermen.

In Rama V's reign, Phuket became the administrative center of a group of tin mining provinces called Monton Phuket, and in 1933, with the change in government from absolute monarchy to a parliamentary system, the island was established as a province by itself.

Report

MT R&D in RDI, NECTEC
~ History, Current and Future of Machine Translation in Thai ~

Information Research and Development Division (RDI) is a unit of National Electronics and Computer Technology Center (NECTEC). NECTEC is a statutory government organization under the National Science and Technology Development Agency (NSTDA), Ministry of Science and Technology. NECTEC was established on September 16, 1986, initially as a project under Ministry of Science, Technology and Energy (the former name of the Ministry of Science and Technology). In December 1991, NECTEC was transformed into a specialized national center under the National Science and Technology Development Agency (NSTDA), a new agency following the Enactment of the Science and Technology Development Act of 1991.

NECTEC has started MT R&D research since 1987 under the Multilingual Machine Translation for Asian Countries Project. It was initiated by CICC (the Center of International Cooperation for Computerization), Japan. The project was the international collaboration between Japan and four other Asian countries, namely, Indonesia, Malaysia, People's Republic of China, and Thailand. The interlingual (IL) approach was considered as an appropriate technique for realizing a multilingual machine translation system efficiently and providing a common research topic for researchers.

At the starting point of the project NECTEC played an importance role as a representative from Thailand and a hub to make collaboration between universities in Thailand. In 1992, NECTEC set up Linguistics and Knowledge Science Laboratory (LINKS) to run NLP research for Thai language including the Multilingual Machine Translation for Asian Countries project. By the end of the project in 1995, NLP resources such as algorithms for Thai language, English to Thai and Thai to English electronic dictionaries, part of speech tag set, analysis rule for Thai as well as Thai corpora had been developed.

In 1996 LINKS and Software Laboratory had been integrated and changed to Software and Language Engineering (SLL). The NLP researches have been continued. The existing algorithms and tools for Thai Language has been modify and increase their efficiency to serve other topics such as search engine for Thai, Thai speech synthesis and Thai to English Machine translation.

Due to strength of collaboration between NECTEC and NEC Corporation, Japan, as partners on the CICC project, NECTEC has started English to Thai Machine Translation with the intensive collaboration with NEC Corporation

In 2000, NECTEC launched a first web-based English to Thai Machine Translation called, "ParSit" It is rule-based Machine Translation with interlingual representation. PARSIT contains around 80,000 vocabularies. It is provided to Thai people with free of charge. ParSit is available at http://www.suparsit.com.

In 2001, NECTEC reorganized its body to ECTI (Electronics, Computing, Telecommunication, and Information) concept.., SLL had been placed to Information concept and changed to Information Research and Development Division (RDI). Text processing section in RDI takes charge of ParSit service, MT research, and NLP infrastructures.

Up to now, ParSit serves about 8,700,000 pageviews. The average amount of IP users are around 1500 per day and the visited pageviews are around 15,000 pages per day. To improve quality of translation service, we developed a server-client based engine to support a large amount of users at the same time. We also applied a proxy-based technique to cache the translation results. This helped users to retrieve translation results more quickly and decrease the bandwidth usage. To improve quality of MT, we are currently developing a Post-edit engine by applying machine learning technique to avoid conflict of rules.

In 2004, ParSit was introduced to Cobra Gold 2004, the joint military exercise between Thailand, USA and other countries, to facilitate communication among Thai and foreign soldiers. From the result of the activity, RDI has extended collaboration with Communications-Electronics Research, Development and Engineering Center (CERDEC) in US military We developed a specific domain, military dictionary, to assist the domain specific translation.

Moreover, mobile applications causes rapid change of the accessing information paradigm. MT in specific platform is also recognized as a research issues. RDI also set up a web-service for SMS translation, which is one of specific environments of MT.

For Thai to English Machine translation, we developed prototype in 2003.It composed of sentence and word segmentation engine, which is suited for translation. Currently, there are around 4000 analysis rules and 20,000 vocabularies. It can be translated a simple sentence pattern and some of compound sentences pattern. The prototype will be available soon.

In the future, speech technology and machine translation, which both available in RDI, will be combined. RDI plans to develop a. speech to speech machine translation, which need a speech synthesis, machine translation and speech recognition as three mains components. In addition, multilingual translation is tended to be a joint project with ASEAN countries, such as Cambodia, Laos, Myanmar and Vietnam. We developed a multiligualization network project and got an endorsement from ASEAN working group committee.

We extend our focus on other MT approaches, statistical based MT and Example based MT. These approaches need a large amount of parallel corpus. We are developing a parallel corpus and necessary tools, such as word alignment, sentence alignment and document alignment.

RDI keeps our interest in MT R&D research and we are very pleased to collaborate with others to realize the machine translation as a real world application.

Thepchai Supnithi, researcher at NECTEC
Monthika Boriboon, researcher at NECTEC
Virach Sornlertlamvanich, TCL director

Report

NLP R&D in Thai Computational Linguistics Laboratory (TCL)

Virach Sornlertlamvanich, Canasai Kruengkrai

What is TCL?

Established in November 2002, Thai Computational Linguistics Laboratory (TCL) is a partnership-laboratory of Computational Linguistics Group (Japan), under Keihanna Human Info-Communication Research Center (KICR) of National Institute of Information and Communications Technology (NICT). TCL is located at Thailand Science Park, in the building of National Electronics and Computer Technology Center, where many research and development on info-communications, material, and bio-technologies being conducted.

It is commonly known that many Asian languages have their own unique problems. Existing methods in Natural Language Processing (NLP) may not be practical when they are directly applied to Asian languages. Customizing existing methods and developing new technologies have been TCL's main research activities over the past three years.

Research Areas

Human Language Technology (HLT)

The goal of HLT is to enable computers to interact with humans using natural language capabilities. Research on HLT includes Language Resource Management, Language Processing, Language Generation, and other related technologies such as probabilistic parsing, word sense disambiguation, etc.

Intelligent Information Infrastructure (III)

R&D on III is the attempt to study and develop general equipments for retrieving and distributing enormous information in both statistic and dynamic manners. Research topics include Information Retrieval, Information Extraction, Data Mining, and Semantic Web. The objective of TCL on III is to carry out the information and knowledge development to bridge the digital divide.

Open Source Software Technology (OST)

Open Source has a unique nature in the software development in terms of revealing the software source code to the public eyes. This means that through the Internet, any software developers are able to collaborate in order to strengthen the capability of the free software without the obstruction of the license agreement. Many governments are taking the OSS as a policy to meet the goal of cost reduction and self-reliance in the software development. TCL intends to develop common software, based on the achievements in NLP research, according to the OSS philosophy.

Current Projects and Collaborations

TCL's Computational Lexicon

At TCL, an initial effort has been made to develop a lexical database named the TCL's Computational Lexicon that aims to serve as the fundamental linguistic resource for Thai NLP research. We design both terminology and ontology for structuring the TCL's Computational Lexicon based on the concept of computability and reusability, since we discovered that word sense representation in general dictionaries with the descriptive manner is not suitable for comparing or distinguishing by computers.

The TCL's Computational Lexicon has three levels of information, including morphological, syntactic, and semantic, and systematically discriminates word sense using a set of logical and semantic constraints. The logical constraints are capable of dealing with the absence of relatedness of word meanings. The semantic constraints try to discover preferences of syntactic arguments of thematic roles. In addition to the web-based editor, the TCL's Computational Lexicon offers lexicographers with statistical corpus-based tools for inserting, updating, and refining lexical entries.

Automatic Language Identifier

Although over 6,000 languages are currently spoken in the world, only a small number of them has been appropriately represented on the Internet. A collaborative project among several research communities named the Language Observatory Project has been undertaken to raise public awareness on this issue, and encourage support to the processing of those languages now falling through the Internet.

TCL provides a mechanism that can automatically identify the language of a given text directly, regardless of its coding system (such as ISO-8859-1), by considering the text in a more fine-grained encoding as the string of bytes. We have developed learning algorithms based on string kernels that can efficiently compute the similarity between two texts and accelerate the kernel computation with a data structure called suffix trees. The current software module can identify more than 20 languages.

Multi-lingual Search Engine

An Internet search engine enables people to find information on the World Wide Web. However, the performance of dictionary-based search engines is directly affected by the accuracy of word segmentation algorithms. To overcome this problem, TCL proposes a search method that does not rely on word segmentation algorithms. The data is considered to be the sequence of characters and indexed character by character. The enhanced suffix array is used for indexing the data.

The advantage of this indexing method is that it guarantees all search strings to be found, whereas the word indexing method depends on the word segmentation. This indexing method can also be applied to other languages, since it does not require any dictionary and language-specific knowledge.

Knowledge Unifying Initiator (KUI)

Under a joint research project, Intercultural Collaboration Experiments (ICE 2005), conducted by Kyoto

University, TCL had an opportunity to participate in the project and developed a prototype software called Knowledge Unifying Initiator (KUI), which is a web-based application that integrates the chat ability to support a collaborative task via the Internet. KUI provides a set of topics for participants to share their opinions, knowledge, and expertise. Participants can exchange ideas and discuss with others to unify their knowledge about the given topic.

The prototype software was successfully tested with translating the medical questioning (originally written in English) into 4 languages, including Thai, Chinese, Japanese, and Korean. The list of questions is the diagnostic interview for deciding medical treatment usually asked when a patient is first admitted to a hospital. We intend to enhance and generalize the capability of KUI to work with other tasks, such as multilingual dictionary construction, law discussion, etc.

Report

Machine Translation R&D in Malaysia – A Brief Update

Normaziah Abdul Aziz

Introduction

Machine Translation as part of the Natural Language Processing R&D in Malaysia begun actively in the 80's and has a history for itself. During the period of 1988 to 1994 works on Natural Language Processing inclusive MT were at its peak with 60 researchers with several product components in various institutions i.e. University Science Malaysia (USM), University of Technology Malaysia (UTM), National University of Malaysia (UKM) and Dewan Bahasa dan Pustaka (DBP), an agency that takes charge on issues related to Malay language. However works in this area declined for several reasons: i) lack of formal linguistics to work with the technical team, ii) lack of financial support (due to low appreciation by the sponsors), iii) low commercialization potential (due to high expectations, e.g. in translation quality); and iv) failure in deployment for either social or commercial purpose. In short, these R&D efforts were unable to make big impact to the country.

A "retrospection" discussion was made by a group of MT and NLP players and we realized that the our local MT and NLP related R&D works have limited resources and were un-orchestrated. Hence, a critical need to focus and align MT and NLP research efforts, optimize resource allocation and to collaborate to achieve greater results.

New R&D working mode

MIMOS, a prominent Malaysian R&D institution on ICT, has proactively driven the development of a research cluster on MT and NLP related under the name of Language Technology Research Cluster. The modus operandi is to develop a technology roadmap for Language Technology and use it as the common agenda to rally and develop the research community around it. Since the roadmap development is consensus driven, the very process of developing the roadmap helps to seed the cluster. Participation from the computer scientists, speech engineers, computational linguists and linguists are very encouraging. These research cluster members are from various institutions – the local universities (USM, UTM, UKM, UM, UPM), DBP, the National Institution of Translation, Malaysian Translation Association, Malaysian Linguistic Association, related industry players and MIMOS. All participants are very supportive of pulling their isolated works together into a coherent whole. Now, the current NLP players work together under the umbrella of the Language Technology Research Cluster. Machine Translation R&D is no exception in this new exercise.

Current Status of MT in Malaysia

MIMOS and University Science Malaysia (USM) have worked on an online English-Malay (and vice-versa) machine translation to address the issue of language barrier among the digital divide community. The system is

an example-based with synchronous structured string-tree correspondence (S-SSTC) MT. To date, it focuses on agriculture and health domain in response to the needs of the digital divide community to understand useful information from the websites. The system has been deployed for usage since August 2004 at www.terjemah.net.my and has translated an average of 14,300 web pages per month. The translation quality improvement is an ongoing R&D effort that both MIMOS and USM are attending to, currently.

The above project is planned for further expansion in terms of MT techniques, translation domains, language pairs, and audience. On the same token, new MT projects with different focuses or approaches are encouraged and will complement existing MT R&D in Malaysia. Automatic translation will not only be for main languages in Malaysia i.e. Malay, Chinese, Indian and English but effort is planned to apply MT for the minority local native languages of the Iban, Kadazan and Bajau communities, among others. In other words, protection and projection of languages in Malaysia is one of the Language Technology research cluster's initiatives.

Major concern to support the country's translation industry via technology (such as Machine Translation, Computer Aided Translation Toolkit and Translation Memory) is also being addressed. This is through various healthy collaboration and strategic planning among the R&D institutions, universities, the National Institute of Translation, the Malaysian Translation Association, and the stakeholders.

Way Forward

With the Language Technology Research Cluster in action and the Language Technology Roadmap in placed (and constantly revised), the Machine Translation R&D which is one of the major components in the roadmap, will chart into a more dynamic R&D mode and produce products that benefit the society.

Report

State of the Art of Machine Translation in Vietnam

Dinh Dien, Hoang Kiem

In Vietnam, the machine translation has attracted the attention of the Vietnamese linguists (e.g. Nguyen Ham Duong, Nguyen Duc Dan [1]) since the '70s. The first English-Vietnamese machine translation system was effected by Logos of the USA under the sponsorship of the US Air Force in the early years of the '70s. However, regrettably, this program was cancelled shortly after the end of the Vietnam War (1975) [2]. In the late years of the '80s, several IT companies (e.g. SoftTech, Seatic, ...) and research groups of the universities (e.g. Polytechnique university of HCMCity, Natural Sciences University of HCMCity, ...) started to invest in the research and construction of the English-Vietnamese machine translation system. In the early years of '90s, one of those companies named SoftTech announced the first version of English-to-Vietnamese machine translation titled as EVTran [3]. In 1998, realizing the ever-increasing demand of machine translation in Vietnam, the national IT-steering board organized a tender aiming at selecting the feasible machine-translation software to be invested for further improvement and perfection. The final ranking was: The 1st prize to EVT software of the machine translation research of the IT faculty of the University of Natural Sciences, HCM City; the 2nd prize to the Ban-Mai company and the 3rd prize to EVTran of SoftTech. In 1999, HCM City financed the EVT software and before the commissioning, this software was appraised as highly reliable by the English teachers based on the translation of various sentences extracted from different sources, of various categories at different levels and by the IT magazine PCWorld of Vietnam. The announced result (PCWorld Vietnam, issue 6/1999) shows that the translation exactness of EVT is 65% for easy sentences, 50% for average sentences and 35% for difficult sentences.

Up to now, there have been more efforts in the machine translation in Vietnam with new groups [4], [5] (e.g. Polytechnique university of Danang, IT Faculty of Technology University of Vietnam National University of Hanoi, Polytechnique university of Hanoi, Institute of IT in Hanoi, etc.) and in which there have been several groups proceeding with the Vietnamese-to-English machine translation with still limited outcome though due to the absence of the indispensable preliminary background for the analysis of the Vietnamese. Besides, there is also the website (http://www.latl.unige.ch/vietnamese) of English-to-Vietnamese and French-to-Vietnamese translation of Doan Nguyen Hai. As far as methods and models are concerned, the majority of the machine translation systems (EVTRAN of SoftTech, EVT of the University of Natural Sciences of HCM City, etc) related to the translation of Vietnamese have based themselves on the approach of RBMT. Recently, there have been the trends shifting to the approach of statistical MT (SMT) and corpus-based MT (CBMT) or machine learning.

Most recent is a project of constructing "an MT archive of languages" at the website www.MT-Archive.info. This electronic repository (and bibliography) of articles, books and papers in the field of machine translation and computer-based translation technology is compiled by John Hutchins of the European Association for Machine Translation on behalf of the International Association for Machine Translation. It is hosted on a

website at the Information Sciences Institute (University of Southern California). In which the latest updating (July, 2005) on the Vietnamese-related machine translation (for both English-to-Vietnamese and Vietnamese-to-English translation) includes 4 papers presented at the reputable conferences on machine translation in the world. The remarkable paper is [6], in which the authors present the translation model of BTL (Bitext Transfer Learning) to learn the transfer rules from the bilingual English-Vietnamese corpora instead of the rules created by human beings formerly

Remark: As far as linguistics is concerned, as English belongs to the inflection language whereas Vietnamese to the isolation one, there are a lot of differences in morphology, syntax and semantics between these 2 languages. For example, the identification of word boundary in Vietnamese is not simply limited to the use of blanks as in English; the word order in the semantic level is also different from that of English (in Vietnamese, nouns stand before adjectives, whereas, it is just the opposite in English). As far as lexicalization is concerned, etc... Therefore, in the course of English-Vietnamese machine translation, the IT specialists should make use of the research results of the comparative linguists for the comparison of the similarities and differences between English and Vietnamese.

References:

[1] Nguyễn Đức Dân (1973), "Dịch máy như thế nào ?", Tạp chí Ngôn ngữ, Viện Ngôn ngữ học Việt Nam, số 1, tr.55-60 (translated: *"How MT works ?", Journal of Language, Linguistic Institute of Vietnam, No.1*) .

[2] Hutchins J. (1985), *Machine Translation: Past, Present,Future*, Ellis Horwood Ltd. Publisher.

[3] Lê Khánh Hùng (1991), *Hệ dịch tự động Anh-Việt*, báo cáo đề tài cấp bộ, Viện CNTT, Hà Nội (translated: *"The Automatic English-to-Vietnamese MT system", Report of National Project, IT Institute*) .

[4] Phan Thị Tươi (2001), *Một số kết quả về dịch tự động Anh-Việt*, báo cáo đề tài cấp Thành phố, ĐHBK-TP HCM (translated: *"Some Results of Automatic English-to-Vietnamese MT", Report of HCM City Project, Polytechnique university of HCMCity*).

[5] Lê Anh Cường, Phạm Hồng Nguyên, Hồ Sĩ Đàm, Nguyễn Phương Thái, Nguyễn Văn Vinh (2002), "Phương pháp chọn nghĩa dịch dựa vào thống kê ngôn ngữ đích trong một hệ thống dịch tự động Anh-Việt", *báo cáo Hội thảo "một số vấn đề chọn lọc của CNTT"*, Nha-Trang, 6/2002, tr.7 (translated: *"Methods for choosing meaning based on statistics of target language in an automatic English-to-Vietnamese MT system", Report at Seminar on some issues of IT, NhaTrang, Vietnam*).

[6] Dinh Dien, Kiem Hoang, & Eduard Hovy (2003), BTL: a hybrid model for English-Vietnamese machine translation *MT Summit IX*, New Orleans, USA, 23-27 September 2003 [PDF, 199KB].

Report

Machine Translation Activities in India

Om Vikas

India is a multi-lingual multi-script democratic country with 22 constitutionally recognized languages and 10 Indic scripts in vogue. Less than 5 percent of India's population, that is 1.08 Billion, can work in English. There are 28 states and 7 union territories. Communication within the state is in the state official language; communication in the central government is bilingual in English. Communication within the state is in the state official language; communication in the central government is bilingual in English and Hindi written in Devanagari Script.

Linguistic cultural similarity threads the multilingualism in the country. Indic scripts are phonetic based with alphabetic similarity in vowels and consonants, vowel-modifier in CV, CCV, CCCV syllables. They have similar script grammar. Indian language sentences follow SOV word order. However, word order is relatively free. We may group them as Indo-Aryan family (Hindi, Marathi, Konkani, Sanskrit, Punjabi, Gujarati, Bengali, Assamese, Manipuri, Maithili, Dogri, Bodo, Santhali and Oriya) in North India; Dravidian family (Tamil, Telugu, Kannada and Malayalam), and Perso-Arabic family (Urdu, Sindhi, Kashmiri).

ICT is an enabling technology yielding higher productivity, better quality, and universal access to information. E-governance aims at ensuring transparency and people's participation in socio-economic development. But English-centric ICT leads to sprawling digital divide. With this backdrop Government of India launched a focused program on Technology Development for Indian Languages (TDIL) in 1990. TDIL became a Mission-mode program in 2000 with 13 resource centers covering all Indian languages. In 2005, mission-mode projects have been identified. Request for proposals will be issued to develop language technology products/services/resources with Public-Private partnership.

Machine Translation activities, began as R&D projects in 1986 at IIT Kanpur. Rajeev Sangal formulated the first project on "Machine Translation among Indian Languages" which was supported by Government in 1986, 1992, and 1995. Core example based "Anusaarak" technology was developed. This has limited sense disambiguation. This is simple MT System between Indian Languages: Kannada-Hindi, Marathi-Hindi, Punjabi-Hindi, Telugu-Hindi and Bengali-Hindi. **(http://anusaarak.iiit.net)**

RMK Sinha at IIT Kanpur formulated the second project on Machine Aided Translation between English and Hindi. It was supported by the Government in 1995, 2000, & 2003. This resulted into "AnglaBharati" technology – rules plus example based MT engine – that was demonstrated for English to Hindi with 60000-root words lexical database in officialese, health, and agricultural domains. This is being adapted for other Indian languages, e.g. Marathi & Konkani, Assamese & Manipuri, Bangla, Oriya, Sanskrit and Tamil. On-line MT System is available on **http://anglahindi.iitk.ac.in/**

Third project "MaTra" on translation of News summaries and News stories in general domain was supported at NCST Mumbai in 1995 and 1999. This uses Frame based approach and involves user intervention to capture context. The MaTra prototype, Vaakya system, was developed at NCST for web based translation service for English news stories into Hindi.

The Department of Official Language, Government of India at CDAC, Pune, funded fourth project on Machine Translation from English to Hindi in officialese domain. H. Darbari formulated this. This is known as MANTRA. This uses Tree Adjoining Grammar and lexical tree of SL into lexical tree of TL transfer approach.

Fifth project on Machine Translation from Hindi to English was funded by Dept of Information Technology, Government of India at IIT Kanpur in Feb 2002. This is based on Example Based Machine Translation (EBMT) paradigm. Hindi conversation includes sometimes English words also. These words written in Devanagari needs to be recognized.

"Shakti" MT System at IIIT Hyderabad uses EBMT model of Carnegie Mellon University. This is between English-Hindi, English-Marathi and English-Telugu **(http://shakti.iiit.net).** "Shiv" MT System (from English to Hindi & Hindi to English) is being developed at IIIT Hyderabad. This also uses EBMT paradigm. **(http://shiva.iiit.net)**

UNL based MT System supports 15 world languages. Pushpak Bhattacharya at IIT Bombay initiated, develop Universal Word Dictionary, Enconverter and Deconverter for Hindi in 1996.

In order to build up language database for MT, 1 Million pages parallel corpora project in 11 languages was funded at C-DAC, Noida. This is available for researchers to use. Efforts are made to establish Linguistic Resource Center in Collaboration with LDC & ELRA.

Government plans to launch fonts and basic software tools for Indian languages for free use. For Tamil and Hindi, these have been launched, launch of fonts and software tools for other will be made in phased manner. Simple Machine Translation System is also included in the tool-kit.

Mission Mode Projects: Telephone based Information Access and Cross Lingual Information Retrieval include MT component. Technology components identified for development include POS Tagger, Morph Analyzer & Generator, Chunking engine, Bilingual dictionary, Summarization engine, Transfer engine, Annotated corpora, Named entity recognizer. Speech-to-Speech Translation is the ultimate objective.

Reference:

http://tdil.mit.gov.in

http://www.ildc.gov.in

Report

Machine Translation R&D Activities at IIT Kanpur

R. M. K. Sinha

The work on machine translation at IIT Kanpur started in early eighties when we proposed using Sanskrit as interlingua for translation to and from Indian languages (Sinha 1984, 1989). As English continues to be the real language barrier in the Indian society, a methodology for translating English to all Indian languages, named AnglaBharti technology, was developed and launched in 1991. AnglaBharti is a pattern directed rule based system with context free grammar like structure for English (source language). It generates a `pseudo-target' (Pseudo-Interlingua called PLIL) applicable to a group of Indian languages (target languages) such as Indo-Aryan family (Hindi, Bangla, Asamiya, Punjabi, Marathi, Oriya, Gujarati etc.), Dravidian family (Tamil, Telugu, Kannada & Malayalam) and others. Within each group the languages exhibit a high degree of structural homogeneity. An attempt is made to resolve most of the ambiguities using ontology, syntactic & semantic tags and some pragmatic rules. The unresolved ambiguities are left for human post-editing. The first prototype based on this methodology was built for English to Tamil in 1991. Later a more comprehensive system, named AnglaHindi, was built for English to Hindi translation. This technology has been transferred on a non-exclusive basis to ER&DCI/CDAC Noida for commercialization. It has also been transferred to eight different organizations for development of MAT systems for English to different Indian languages catering to 12 regional languages of the country (IIT Mumbai: Marathi & Konkani; IIT Gwahauti: Asamiya & Manipuri; CDAC Kolkata: Bangala; CDAC(GIST group) Pune: Urdu, Sindhi & Kashmiri; CDAC Thiruananthpuram: Malayalam; TIET Patiala: Punjabi; JNU New Delhi: Sanskrit; and Utkal University Bhuvaneshwar: Oriya). The AnglaBharti system architecture has undergone considerable modification in AnglaBharti-II (2004). AnglaBharti-II is hybridization of RBMT and EBMT. During the development phase, when it is found that the modification in the rule-base is difficult and may result in unpredictable results, the example-base is grown interactively by augmenting it. At the time of actual usage, the system first attempts a match in example-base before invoking the rule-base. We made provision for automated pre-editing & paraphrasing, generalized & conditional multi-word expressions, recognition of named-entities, domain customization tools and incorporated an error-analysis module and statistical language-model for automated post-editing. The purpose of automatic pre-editing module is to transform/paraphrase the input sentence to a form which is more easily translatable. The entire system is pipelined with various sub-modules. All these have contributed significantly to greater accuracy and robustness to the system. A random testing on the system yielded BLEU and NIST scores as 3.41-4.88 and 0.15-0.39 respectively using two reference translations.

In 1995, we developed another MT methodology, named *AnuBharti*, based on EBMT paradigm. The translation is obtained by matching the input sentence with the minimum 'distance' example sentence. We store the examples in generalized form to contain the category/class information to a great extent. This makes the example-base smaller in size and further partitioning reduces the search space. The creation and growth

of the example-base is also done in an interactive way. This methodology has been used for Hindi to English translation. The Anubharti approach works more efficiently for similar languages such as among Indian languages. In such cases the word-order remains the same and one need not have pointers to establish correspondences. The strategy has now been generalized in AnuBharti-II (2004) to cater to Hindi as source language for translation to any other language, though the generalization of the example-base is dependent upon the target language. The core of AnuBharti-II architecture is a generalized hierarchical example-base. Hindi like all other Indian languages is a relatively free word-group order language. The input Hindi sentence is converted into a standardized form to take care of word-order variations. This requires a shallow grammatical analysis of Hindi. This makes the paradigm used in AnuBharti-II a hybrid paradigm. Human post-editing is performed primarily to introduce determiners that are either not present or difficult to estimate in Hindi.

Besides these, we have also developed a translation system for bi-lingual text in Hinglish (Hindi mixed with English) and working on system for speech to speech translation.

Bibliography

http://www.cse.iitk.ac.in/users/rmk/proj/proj.html#mt

Sinha R.M.K. and A. Thakur. 2005a. Machine Translation of Bi-lingual Hindi-English (Hinglish) Text, *10th Machine Translation summit* (MT Summit X), Phuket, Thailand.

Sinha R.M.K. 2005b. Integrating CAT and MT in AnglaBharti-II Architecture, *EAMT 2005*, Budapest, Hungary.

Sinha R.M.K. 2004. An Engineering Perspective of Machine Translation: AnglaBharti-II and AnuBharti-II Architectures, *Proceedings of International Symposium on Machine Translation, NLP and TSS (iSTRANS-2004),* Tata McGraw Hill, New Delhi.

Sinha R.M.K. and others. 1995. ANGLABHARTI: A Multi-lingual Machine Aided Translation Project on Translation from English to Hindi, *1995 IEEE International Conference on Systems, Man and Cybernetics*, Vancouver, Canada, pp 1609-1614.

Sinha R.M.K. 1989. A Sanskrit based Word-expert model for machine translation among Indian languages, *Proceedings of workshop on Computer Processing of Asian Languages*, Asian Institute of Technology, Bangkok, Thailand, pp 82-91.

Sinha R.M.K. 1984. Computer processing of Indian languages and scripts - Potentialities and Problems, *Jour. of Inst. Electron. & Telecom. Engineers*, vol.30,no.6, pp 133-149.

Report

Natural Language Processing Activities at Indian Institute of Technology Bombay, India

Natural Language Processing (NLP) Activities at IIT Bombay started in 1996 with the ***Universal Networking Language (UNL)*** project funded by the United Nations University, Tokyo. 15 research groups across various nations participated at that time in this UN initiated international effort. UNL is a sentential meaning representation scheme which essentially represents the knowledge in the form of binary predicates, $r(u_1,u_2)$, where r is a semantic relation and u_1 and u_2 are disambiguated words expressed using the form *ew(restr)* where *ew* is typically an English word and *restr* is the meaning constraining expression (*e.g., dog(is-a>mammal)*). There is a huge repository of such disambiguated concepts to which corresponding language strings are linked along with syntactico-semantic attributes. Creating and maintaining such a large lexicon is a major task of the UNL based language processing.

During our course of work with UNL at IIT Bombay, we quickly realized that *word-knowledge* is the heart of the system. Since our work was in the trilingual setting of *English, Hindi* and *Marathi* (a language of western India) we set to creating ***language resources and processing tools of for Hindi and Marathi***. A major effort undertaken was the creation of ***Hindi and Marathi Wordnets.*** We also saw that the verb knowledge base is not deep enough in most ontologies existing in the world. This prompted us to working on English and Hindi ***Verb Knowledge Bases.***

Annotated corpora are very valuable and scarce resources in any language. This is needed for most statistical language processing systems. However, many of the disadvantages of resource poor languages can be offset if the language in question is rich in morphology. A major work going on at IIT Bombay is the creation of ***Morphology Analysers and Part of Speech Taggers for Hindi and Marathi*** both of which have rich morphological structures. Our approach is based on *bootstrapping on a small corpus tagged by* ***Rule Based Tagger*** and then applying statistical techniques to train a machine. Currently HMM and Conditional Random Field based approaches are being examined.

Since the above activities are highly semantics oriented, ***Word Sense Disambiguation*** assumes a very important role in our activities. We have introduced the concept of *soft word sense disambiguation. word sense disambiguation* which ranks the senses according a probabilistic estimate. A major application of the Hindi wordnet has been in ***Hindi Word Sense Disambiguation*** using a sense intersection approach.

We have done extensive work on ***Meaning Based, Multilingual Search Engine.*** The search engine is employed in the agricultural domain and is called **AgroExplorer.** It uses internally the UNL representation of documents, and queries too are in UNL forms. The search can fall back from meaning based search to concept based search to keyword based search. As a part of the activities we have worked on ***Multilingual Keyword Search*** and ***Similarity Based Retrieval*** which need NLP techniques.

Summarizing, therefore:

List of NLP activities at IIT Bombay

A. Interlingua (UNL) Based Machine Translation in the Trilingual setting of Hindi, Marathi and English
B. Creation of Lexical Resources: Hindi and Marathi Wordnets, Hindi and English Verb Knowledge Bases, Lexicons with Elaborate Syntactic and Semantic Attributes
C. Tools for Indian Language Processing: Morpohology Analyser and POS Tagger for Hindi and Marathi
D. UNL Based Text Classification, Clustering and Summarization (for English)
E. Word Sense Disambiguation (Hindi and English)
F. Meaning Based, Multilingual Search Engine (Trilingual currently)
G. Marathi Spell Checker
H. Hindi and Marathi Corpus Search Utility: *Concordancer*

Accomplishments

a. Hindi wordnet has 16546 synsets (corresponds to 35650 unique words).
b. Marathi wordnet: 9625 synsets (corresponds to 15269 unique words).
c. Universal Word-Hindi Dictionary has 101438 entries.
d. English Verb Knowledge Base (VKB hierarchy) has 6350 nodes and the corresponding Hindi VKB has 3126 nodes.
e. Hindi POS Tagger shows good performance on BBC Hindi web pages.
f. Marathi Morph Analyser and POS tagger (for verb) and Hindi Morph Analyser ready.
g. 4-Stage Meaning Based, Multilingual Search Engine is working on Agricultural Corpora.
h. Hindi Word Sense Disambiguator (Lesk algo based) works with average accuracy of 40%.
i. Soft Word Sense Disambiguator (English) attains about 75% accuracy (first two senses).

Websites

www.cfilt.iitb.ac.in (for resources and tools)

http://www.cse.iitb.ac.in/~pb/pubs-yearwise.html (for publications)

Report

NLP Research Activities at IIT Kharagpur

Monojit Choudhury, Sudeshna Sarkar, Anupam Basu

The overall goal of the NLP research activities being pursued at the Communication Empowerment Laboratory (CEL), IIT Kharagpur, is to develop complete NLP tool sets and lexical resources for Bengali, leading to applications such as Text-to-speech (TTS) and speech-to-text systems, Machine Translation (MT) systems from Bengali to English and other Indian languages and vice versa, and visual to natural language converters for Augmentative and Alternative Communication (AAC) systems. Since the structure and the vocabulary of Hindi are quite similar to that of Bengali, similar R&D activities for Hindi are also being carried out. Porting of NLP applications on embedded devices and study of their dynamic execution characteristics for design and development of optimized NLP architectures are the goals of the embedded system research group at CEL. Yet another parallel thread of research, very unique to CEL, is exploration of synthetic and analytical computational models for language change and their possible applications in NLP.

Apart from the existing mono-lingual corpora for Bengali and Hindi developed by Central Institute of Indian Languages, which are being used for NLP research, CEL has also developed its own domain specific corpora for these languages by gathering documents from the web. A POS-tagged corpus for Bengali is being developed semi-automatically. Currently, the corpus consists of 30000 tagged words. The tag set used for annotation, which consists of forty tags, has been designed at CEL. A root word lexicon with POS information and a pronouncing dictionary for Bengali each with 100000 entries have been developed in-house to facilitate POS-tagging and text-to-speech synthesis.

We are working towards a verb-centric semantic net-like interlingua based MT system. The interlingua has been designed keeping in mind the structure of the source and the target languages. Tools for analysis of the source language (Hindi and Bengali) text that have been developed include FST-based morphological analyzers for inflectional morphology, an HMM based POS tagger for Bengali, rule-based POS-tagger and chunker for Hindi, and a novel affinity-based greedy algorithm for Bengali chunker. Presently, we are working towards sentence-level parsers for both the languages. A language-independent sentence generation engine has been developed that takes the grammar of a language as input and generates sentences in the target language from the interlingua representation. Grammars for Bengali and English are being developed to facilitate natural language generation in these languages.

Shruti – a partnym-based concatenative speech synthesizer has been developed at CEL, which has support for both Bengali and Hindi. A rule-based grapheme-to-phoneme (G2P) converter for Hindi and a hybrid G2P converter for Bengali have been built to handle pronunciation changes. *Shruti* has been ported to low end PDAs and other handheld devices. At present we are working on the intonation, prosody and duration for these

languages to yield more natural speech. Work on automatic speech recognition has already been initiated with a considerable success in limited vocabulary connected word recognition. We are interested in developing large vocabulary speaker-independent continuous speech recognition system for Bengali.

Sanyog – a multi-modal communication system for people with neuro-motor and/or speech impairments is an NLP based application that has facilities for converting sequence of icons to natural language sentences, with speech output (from *Shruti)* and an editor with character and word level predictions. Currently, *Sanyog* has support for Bengali, Hindi and English. A lightweight editor for Bengali and Hindi with spell-checking facility has been developed that supports several fonts, encoding schemes like Unicode and ISCII, and different input schemes like ITRANS, wx notation etc. It also supports conversion between different encodings and fonts. The isolated word spell-checker has a robust phonetic error detection and correction module. Presently, we are working towards context-dependent word error detection and correction for both typed and OCR generated documents.

In the domain of diachronic linguistics, we have developed a multi-objective optimization model for emergence of the schwa deletion pattern in Hindi, which has been used to design a rule-based schwa deletion algorithm. Multi-agent based simulation models for the same phenomenon are also being explored. Currently, we are trying to model the morpho-phonological changes that have affected the Bengali verbal inflections. This line of research helps us understand the dynamics of language change and its probable application to NLP.

Detailed information on our activities and a complete list of publications in related fields are available at *http://cel.iitkgp.ernet.in*

Report

NLP Activities at the Department of Computer and Information Sciences, University of Hyderabad, India

The Department of CIS, University of Hyderabad, India, has been engaged in Language and Speech Technology Research for over 15 years now. A wide variety of topics have been taken up with primary focus on English and Indian Languages.

1. Syntactic Analysis: While a lot of progress has been made in terms of linguistic grammar formalisms, developing practical descriptive grammars for automatic parsing has not been easy at all. Developing wide-coverage, robust descriptive grammars takes a lot of time and effort and the required linguistic expertise is also not always available. On the other hand, statistical approaches to grammar discovery and parsing look attractive but are practicable only when sizable parsed corpora are available for training. Given this scenario, we have initiated a unique architecture called UCSG architecture that combines the best of both worlds. Grammars and parsing systems have been developed for English and other Indian languages within the UCSG system. The current focus is on wide coverage, robust shallow parsing. A Finite State Grammar is developed from linguistic insights and a large text corpus is chunked using this grammar. Using this as training data, HMMs are built for each type of chunk to rate and rank the chunks obtained by the Finite State Parser. Note that only evaluation is involved and not Viterby search. This is used for local optimization and an A* search is performed to obtain global best chunking for a whole sentence. Thus the parser gives the best parse first while retaining its ability to produce all possible parse outputs. We can bootstrap and re-train the HMMs using the top few parses. Preliminary results show that the approach holds promise - wide coverage, robust shallow parsers can be developed quickly without the need for a parsed training corpus to start with. The method should also be useful for Indian languages where plain text corpora are available but parsed corpora are not yet available [1,2].

2. Word Sense Disambiguation: Words have multiple senses and WSD deals with the task of identifying the correct sense of a word in context. The definition of context is a tricky issue and not all words in a sentence are useful for disambiguating the senses of a given target word. In fact we argue that some words are noise and only interfere with the task. We have shown that by eliminating noise, performance actually improves. In a recent experiment, we have taken syntactically related words as possibly relevant and other words as noise. CMU's Link parser was used to identify syntactically related words. It has been shown that 5 to 12% improvement in performance can be obtained when tested on standard test data compared to best published results on the same data [3].

3. Language Identification: There are one Billion people in India speaking about 150 different languages (not dialects). Twenty two of them have been given constitutional status and are considered to be major languages. These major languages are written in ten different scripts. Many documents are required to be bilingual or trilingual. Mixing up of different languages is a regular and frequent phenomenon. The correspondence

between languages and scripts is not strictly one to one - a script is used to write several languages and a given language may be written in more than one script. ISCII, the Indian standard Script Code for Information Interchange, exploits the phonetic nature of the scripts and assigns common code space to all the ten scripts. Thus the characters that make up a text indicate sounds and not the script or the language. A high performance system has been developed for language identification from small text samples. Multiple Linear Regression has been cast as a Two-class Classification model and n-gram features based on syllabic structure of Indian scripts have been used [4].

4. Text Categorization: A system has been developed for Automatic Categorization of Telugu News Articles using the Bayesian Learning approach. Suitably weighted and normalized tf-idf features have been used. Category-wise inverse frequency has been found to be more appropriate [5].

5. Other Tools: DRISHTI - an OCR engine for Telugu and other Indian scripts, AKSHARA - an advanced multi-lingual Text Processing Environment, Machine Aided Translation System, Spell Checkers, Morphological Analyzers and Generators, Electronic Dictionaries and Thesauri, Corpora (38 Million Words) for Telugu and other Indian Languages. Recent work includes Speaker Independent Continuous Speech Recognition system for Telugu and a Speaker Recognition System [6].

References:

1. G. Bharadwaja Kumar, K Narayana Murthy, 'Towards a Robust Shallow Parser', SIMPLE 2005, IIT (Kharagpur), India

2. G. Bharadwaja Kumar, K Narayana Murthy, 'UCSG Shallow Parser', Comm. to ICON 2005

3. A Sasi Kanth and K Narayana Murthy, 'Significance of Syntactic Features for Word Sense Disambiguation', Lecture Notes in AI, Vol 3230, pp 340-348, Springer-Verlag, 2004

4. K Narayana Murthy and G Bharadwaja Kumar, 'Language Identification from Small Text Samples', To Appear in Journal of Quantitative Linguistics

5. K. Narayana Murthy, 'Automatic Categorization of Telugu News Articles', Proc. of ICOSAL-6, Hyderabad, 2005

6. www.TDIL.mit.gov.in, www.LanguageTechnologies.ac.in

Report

NLP Activities at TIET, Patiala INDIA

R. K. Sharma

The NLP activities started in Thapar Institute of Engineering and Technology (TIET), Patiala India from April 2000 when Department of Electronics, Government of India sanctioned a major project to us in the form of Resource Centre for Indian Languages Technology Solutions Punjabi. Since then, Institute is contributing in the research and development activities in this field focusing on the technical development for Punjabi language.

A bilingual word processor (Likhari) that supports word processing under the windows environment and allows typing and processing in Punjabi language through the common typewriter keyboard layout has been developed in this Institute. The features incorporated in Likhari include online active keyboard for users who do not know how to type in Punjabi, choice of phonetic and Remington keyboard layouts with composition reference, bilingual spell checker for Punjabi and English, bilingual search and replace facility, support for sorting the text in English and Punjabi as per the language alphabetical order, support for most of the popular Punjabi fonts and keyboard layouts, online technical glossaries, Punjabi thesaurus and support for .iscii, .txt, .doc, .rtf and .html formats.

An off-line OCR system for Punjabi language has also been developed here at TIET, Patiala. It has good recognition accuracy for laser prints and fine quality documents, almost 99%, and for books, photocopied papers and medium degraded documents, it is around 97%. The system supports for almost all non-decorative Gurmukhi fonts. It has an on-screen verifier, inbuilt spell checking facility, automatic skew detection and correction and also upside down image auto detection and correction facility. Presently, the work is in progress to improve the system in order to recognize the degraded Punjabi text. The work is also in progress to develop an on-line OCR system for Punjabi language.

The Resource Centre at TIET, Patiala has developed two ISCII compatible true type fonts, Likhari_P and Likhari_R for Punjabi. These fonts can be used on Windows based platform.

A bilingual spell checker for Punjabi language that supports spell checking facility for Punjabi as well as English under the windows environment has also been developed. The spell checker automatically detects the wrong words and suggests the possible correct spellings. The spell checker successfully solves the problem of time consuming proof reading for the Punjabi text.

Punjabi language is used in both parts of Punjab, in India and Pakistan. In East Punjab (India) Punjabi is written in Gurmukhi script. This is written from left to right. In West Punjab (Pakistan) Punjabi is written in Shahmukhi script. This is written from right to left like Urdu and Persian. A transliteration program has been developed to break the barrier between the Punjabi language written in these two scripts which can convert Gurmukhi Text to Shahmukhi in collaboration with CDAC, Pune.

A technical glossary of around 17,000 English-Punjabi administrative terms has been developed and has also been uploaded on the Internet. An online Punjabi to English dictionary that has about 39,000 Punjabi words

has also been developed. This dictionary has sample sentences for common words and clips of pronunciation are also incorporated in it. The dictionary is accessible in both typewriter and phonetic layout. One can search for a complete string match or a pattern in the dictionary. Sorting algorithm for Punjabi words has been used to sort the Gurmukhi words in the database. The feature of dynamic fonts has also been used in the dictionary. The online Hindi to Punjabi dictionary that has been developed at TIET, Patiala has about 45,000 words. This dictionary displays the words in alphabetical order. The online English to Punjabi dictionary developed here has about 46,000 words. One can search for a complete string match or a pattern in this dictionary also. An online thesaurus for Punjabi language has also been developed. It has about 6,100 words that frequently occur in our day to day life. The thesaurus is accessible in both typewriter and phonetic layout. One can search for a complete string match or a pattern in the thesaurus also.

For the benefit of Punjabis settled abroad and others interested in learning Punjabi through Internet, web-based material for on-line teaching of Punjabi has been developed. This material consists of introduction to Gurmukhi including its orthography. It also contains instructions for drawing alphabets in animation pattern made into Java applets with position of mouth organs while pronouncing a particular Gurmukhi alphabet with audio effects. The other features of online teacher include a vocabulary consisting of pictures and names of living and nonliving things, some common places etc. in English and Punjabi in the form of composition reference. The site also has inbuilt audio effects, description of Punjabi language and Punjabi pronunciation rules etc.

The Resource Centre at TIET, Patiala has also contributed in uploading some literary and religious Punjabi classics. These classics include Bullehshah Diyan Kafian, Farid De Shalok, Heer, Loona, Mirza, Chandi Di Var and Japji Sahib etc. The website contains the detailed description of these classics. Tool tips for difficult words have also been provided in Punjabi. Audio clips of some of these classics have also been included. Some of the Short stories by Amrita Pritam and K S Duggal have also been uploaded. The online contents are available at Center's URL http://punjabirc.tiet.ac.in.

Report

NLP Activities at Computer Science & Engineering Department Jadavpur University, India

Sivaji Bandyopadhyay

Teaching and research activities in various areas of natural language processing like Machine Translation, Text Summarization, Information Retrieval and Named Entity Recognition are going on. NLP activities are currently being pursued for English, Bengali, Telugu and Manipuri. A one semester course on "Natural Language Processing" is being offered for the students of the Master of Computer Science & Engineering course in the department.

Research methodologies being pursued in the different activity areas are as follows:

Machine Translation

The Phrasal Example Based Machine Translation system from English to Bengali identifies the phrases in the input through a shallow analysis, retrieves the target phrases using a Phrasal Example base and finally combines the target language phrases employing some heuristics based on the phrase ordering rules for Bengali. Work has also started for machine translation of Manipuri to English. The Example based Machine Translation System for translating News Headlines from English to Bengali uses a Direct Example base, a Generalized tagged Example base and a Phrasal Example base in the order.

Text Summarization

Work is going on for generating summary from single or multiple documents using deeper approaches. Our approach to generating headline summary from a document set involves breaking up the sentences into a number of smaller text units based on named entities and compiles the units to generate the final summary.

Information Retrieval

We are working for the development of a Natural Language Interface to a Database system following a Template Grammar based approach and the Keyword based approach. The user query is analyzed and translated to a SQL statement that is used to retrieve information from the database. The domain has been chosen as the Indian Railway Information system. Both the systems accept queries in Bengali and Telugu languages.

Language Technology on Bengali

Work has started on building a tagged corpus of Bengali news documents from the archive of a Bengali newspaper. This tagged corpus is currently being used to develop a Bengali lexicon, Morphological Analyzer and a Named Entity recognizer. Named Entity Identification in Indian languages poses an interesting challenge, as there is no concept of capitalization.

Structuring the unstructured data using intelligent extraction is the other area of research. The work in this area is domain specific but has started working on extraction methods across domains. Extractions of information are using linguistic rules and different data mining techniques. The research is underway in different domains and they are Biology (Medline abstracts), Crime, Environmental, Scientific Patent Documents and Resume/posting.

In **Biological domain** the on going research is on a **Relation** that the proteins undergo viz. **Phosphorylation**. It has a Named entity recognizer (M.Narayanaswamy, et. al) and an Acronym Detector for protein terms and they are highly accurate .The relevant reaction is extracted using linguistic rules (Ravi Kumar et. al 2004), (M, Narayanaswamy et. al 2005).

In the **Crime domain** the named entity recognizer uses dictionary and the extraction is using linguistic rules.

In the **Energy domain** the extraction of information is from Journal articles on the web. In this the title, author, address is extracted using HMM and the other information such as methodology is extracted using data mining techniques.

Research is also going on in enabling information in English to be accessed in Indian languages (Sobha, L et. al 2005). Our **Content aggregation** system is deployed in a popular Indian Portal.

References

Arulmozhi, P. Sobha, L, Kumara Shanmugam. B. (2004) " Part of Speech Tagger for Tamil" Symposium on Inidan Morphology, Phonology and Language Engineering, March 19-21, 2004 IIT Khadagpur, 55-57, India

Bhaskaran, S and Vijay-Shanker, K (2003) "Influence of Morphology in Word Sense Disambiguator for Tamil", Recent Advances in Natural Language Processing, Proceedings of the International Conference ICON 2003, 31-39, CIIL, Mysore.

Ravi Kumar, Meenakshi Narayanaswamy and Vijay K. Shanker ," Information Extraction in Biological Domain - Extracting Phosphorylate Interactions", The ACM Symposium on Applied Computing (SAC-2004).

M. Narayanaswamy, K.E. Ravikumar and K. Vijay-Shanker (2005) "Beyond the clause:extraction of phosphorylation information from medline abstracts" Vol 21 Suppl 1 2005, pages i1-i9 doi 10.1093/bioinformatics/btil 011.

Sobha, L and Arulmozhi, P, (2005) "Translingual Information Accessor using Information Extraction-English to Tamil", The 2nd Indian International Conference on Artificial Intelligence (IICAI-05), Pune, India,

Sobha, L. (2003). "Pronominal Resolution In South Dravidian Languages" (Accepted for publication in an edited publication from SALA Organizers), and presented at *23rd* South Asian Language Analysis (SALA 23), University of Texas, Austin,

Viswanathan .S et. al, (2003), "A Tamil Morphological Analyser" , Recent Advances in Natural Language Processing, Proceedings of the International Conference ICON 2003, 31-39, CIIL, Mysore.

Report

NLP Activities at AU-KBC Research Centre

L. Sobha

The Natural Language Processing Division at AU-KBC Research Centre started in 1999 and focuses the research on two major areas: Machine Translation and Intelligent Information Access.

The Machine Translation work at the Center concentrates on Tamil to Indian Languages and English to Tamil translation. As the initial attempt in Tamil to Indian languages translation, we have developed a **Tamil to Hindi Machine Aided Translation System** (MAT). Indian languages are relatively free word order, inflectional and verb final. Any approach, which can exploit these features of Indian languages, is the best suited for translation. The approach envisaged has exploited these features in building the Tamil-Hindi Translation system. This is a robust system, which can handle sentences with one clause embedding. In the course of developing the MAT system we have developed all the pre-processing tools required such as Morphological Analyser for Tamil, Part of Speech Tagger, Noun Phrase Chunker, Word Sense Disambiguater, The Morphological Generator for Hindi, Anaphora Resolution etc and Lexical recourses such as Bilingual Dictionary (Tamil-Hindi -32000 root words). These tools are developed using linguistic rules, statistical methods such as HMM and Machine Learning approaches. The morphological analyzer (Viswanathan, S et. al, 2003) uses Finite State Automata. Here paradigm methodology with FSA is used for handling the morphophonemic changes that occur during suffixation, which is giving wide coverage and accuracy to the morphological analyzer. The POS tagger (Arulmozhi, P et.al 2004) has 12 basic tags and is developed using Rule based and HMM based approaches. Clustering techniques are used in developing the Word Sense disambiguation system (Baskaran. S, 2003) In Anaphora resolution (sobha, L 2003) we identify the antecedent for pronouns using linguistic rules and also using Centring theory. In Lexical resource we have developed the Tamil wordNet that has 50,000 words.

The English to Tamil Translation System is developed for the "Traditional Indian Medicine" domain. It translates the contents in the web pages related to this domain. English being different in structure from Indian Languages we have adopted both rule-based as well as statistical based approaches.

Developing novel methods for retrieving and accessing information from unstructured data are the main thrust research in **Intelligent Information Access**. Here we deal with unstructured text from Tamil and English. The research has brought in very efficient search engines that could retrieve documents relevant to a query using high-level mathematical approaches. It has produced the first Indian language search engine **Kazhugu** for searching Tamil web pages. The query to the engine is in Tamil and a morphological analyser is integrated to the system that enables to search all the inflections that the query word could take and thus enhances the search. This could search the web pages created in thirty encoding schemes.

Structuring the unstructured data using intelligent extraction is the other area of research. The work in this area is domain specific but has started working on extraction methods across domains. Extractions of information are using linguistic rules and different data mining techniques. The research is underway in

different domains and they are Biology (Medline abstracts), Crime, Environmental, Scientific Patent Documents and Resume/posting.

In **Biological domain** the on going research is on a **Relation** that the proteins undergo viz. **Phosphorylation**. It has a Named entity recognizer (M.Narayanaswamy, et. al) and an Acronym Detector for protein terms and they are highly accurate .The relevant reaction is extracted using linguistic rules (Ravi Kumar et. al 2004), (M, Narayanaswamy et. al 2005).

In the **Crime domain** the named entity recognizer uses dictionary and the extraction is using linguistic rules.

In the **Energy domain** the extraction of information is from Journal articles on the web. In this the title, author, address is extracted using HMM and the other information such as methodology is extracted using data mining techniques.

Research is also going on in enabling information in English to be accessed in Indian languages (Sobha, L et. al 2005). Our **Content aggregation** system is deployed in a popular Indian Portal.

References

Arulmozhi, P. Sobha, L, Kumara Shanmugam. B. (2004) " Part of Speech Tagger for Tamil" Symposium on Inidan Morphology, Phonology and Language Engineering, March 19-21, 2004 IIT Khadagpur, 55-57, India

Bhaskaran, S and Vijay-Shanker, K (2003) "Influence of Morphology in Word Sense Disambiguator for Tamil", Recent Advances in Natural Language Processing, Proceedings of the International Conference ICON 2003, 31-39, CIIL, Mysore.

Ravi Kumar, Meenakshi Narayanaswamy and Vijay K. Shanker ," Information Extraction in Biological Domain - Extracting Phosphorylate Interactions", The ACM Symposium on Applied Computing (SAC-2004).

M. Narayanaswamy, K.E. Ravikumar and K. Vijay-Shanker (2005) "Beyond the clause:extraction of phosphorylation information from medline abstracts" Vol 21 Suppl 1 2005, pages i1-i9 doi 10.1093/bioinformatics/btil 011.

Sobha, L and Arulmozhi, P, (2005) "Translingual Information Accessor using Information Extraction-English to Tamil", The 2nd Indian International Conference on Artificial Intelligence (IICAI-05), Pune, India,

Sobha, L. (2003). "Pronominal Resolution In South Dravidian Languages" (Accepted for publication in an edited publication from SALA Organizers), and presented at *23rd* South Asian Language Analysis (SALA 23), University of Texas, Austin,

Viswanathan .S et. al, (2003), "A Tamil Morphological Analyser" , Recent Advances in Natural Language Processing, Proceedings of the International Conference ICON 2003, 31-39, CIIL, Mysore.

Central Institute of Indian Languages

THE GOVERNMENT'S RESOLUTION of the January 18, 1968 on the Language Policy, as adopted by both the Houses of Parliament, emphasized that in the interest of the educational and cultural advancement of the country it was necessary to take concerted measures for the full development of the major languages of India, besides Hindi. The resolution further enjoined upon the Government to prepare and implement a programme in collaboration with the State Governments for the coordinated development of all these languages so that they grow rapidly in richness and become effective means of communicating modern knowledge. This is clearly recognition of the multilingual character of the country. A significant step taken by the Ministry of Education, Government of India to promote Indian languages was the establishment of Central Institute of Indian Languages on July 17, 1969.

The Institute has the responsibility of conducting research in the areas of language analysis, language pedagogy, language technology and language use with a bias towards problem solving and national integration. The major domains in which the Institute works are education, administration, documentation and mass communication. The Institute is involved in the description and codification of smaller languages and in developing models, methods, materials and manpower for their use in education. It is concerned with the status of major Indian languages and the implementation of policies as regards their use as medium of instruction and administration at all levels. This Institute also helps the Government in language planning and lends its assistance in coordinating the development of Indian languages. www.ciil.org

LANGUAGE TECHNOLOGY

Language teaching and learning: The Institute uses language technology in the area of language teaching/learning and in the area of Natural Language Processing. Since 2001, the Institute initiated the process of developing language courses in Bengali (in collaboration with Netaji Subhas Open University, Kolkata, India), Hindi, Kannada, Manipuri, Oriya, Tamil and Urdu to teach them using worldwide web under its On-line Language Teaching programme. The courses in other 3 languages are under preparation and they will be soon available for the general public in the net.

These courses have been developed for those who want to learn them as second languages. These also caters to the needs of the language diaspora who may have a desire to know not only the language structure but also their unique and vibrant tradition, architecture, music, dance, worshiping pattern and constant interaction between speakers of these languages and their worldwide view. As on today, the Bangla, Kannada and Tamil on-line

courses are available at the following urls: www.bangla-online.info, www.kannada-online.info and www.tamil.online.info respectively. It is possible for the learners to pay and register for the courses on-line and obtain credits after completing learning through these courses.

The Institute since 1985 is conducting a Distance Education Course in Kannada for the employees of the Govt. of Karnataka who do not know Kannada. So far nearly 13000 registrants have got the benefit. The learning materials are placed in the internet at http://www.ciil-learnkannada.net

Text Corpora: http://www.corpora.net. The Central Institute of Indian Languages in the past co-ordinated the development of 45 plus million word corpora in Scheduled Languages under the scheme of Technology Development for Indian Languages(TDIL) of the Ministry of Communication and Information Technology.
This corpora was created following sampling methodologies and hence this is a balanced corpora. This is available in Indian Standard Code for Information Interchange or Indian Script Code for Information Interchange(ISCII) format.The Institute also intends to enhance this corpora to the tune of twenty million in each language.
The Institute in collaboration with the Lancaster University has converted the same into UNICODE format. Corpora in this format, in addition to the Lancaster University corpora is also available for users at :http://www.emille.lancs.ac.uk/home.htm

On its own the Institute is now developing corpora in languages recently included into the Eighth Schedule like Bodo, Dogri, Maithili and Santali. All the prose texts available in these languages are keyed since it is not possible to obtain texts from all the sampling domains.Corpora in Indian languages thus developed is maintained and distributed free of cost to the scholars by the Institute for academic purposes.

However, in near future the institute will also be able to provide different kinds of resources in Indian languages against payment/subscription.This activity as is being planned now will form part of the proposal : Linguistic Data Consortium in Indian Languages (LDCIL).

Spoken Corpora: http://www.ciil-spokencorpus.net. The Institute is creating a spoken language corpus in lesser known language of India in collaboration with Uppsala University, Sweden. The aim of this project is to collect, organize and disseminate information on some lesser-known Indian languages, many of which are threatened with extinction. The project will include linguistic documentation (i.e. texts and speech files) as well as documentation anchoring this linguistic material to social and cultural aspects of these communities.

Web-based Translation service Anukriti : www.anukriti.net. The site anukriti.net is launched in collaboration with Sahitya Akademi and National Book Trust. The site has a huge data base on translation, a course on translation studies and an ejournal "Translation Today'. This journal has both web and print versions.

Report

SNLP Activities at C-DAC Noida

V. N. Shukla, Karunesh K. Arora,
Sunita Arora, Vijay Gugnani, S. S. Agrawal

Translation Support System: Translation support System for translation from English to Hindi is based on the ANGLABHARATI technology of Prof. R M K Sinha at the IIT, Kanpur. The system uses hybrid approach of Rule based and Example based approaches and lexicon is general-purpose. However, the system has been applied mainly in the domain of public health campaign. The software is primarily developed at IIT Kanpur in collaboration with CDAC, Noida

Test Bed for Evaluation of English-Hindi Machine Translation System: Researchers, Developers and Investors community would like to measure the performance of the system to test a particular approach, linguistic coverage and to check the adverse effect caused by insertion of a new rule. Translators and post-editors measure the performance based on capability improvement, consistency in translation and the saving in manual efforts. Test Bed is an evaluation framework which consists of Test suite & Evaluation criteria.

Criteria for development of Test Suite -

- Identification of different lexical and structural categories
- Collection of sentences covering variety of lexical and structural components
- Human translation of each sentence by three different translators

To measure closeness of translation

- Objective queries associated with each sentence have been framed which are asked by the evaluator and used to calculate the score to rank the system in the scale of 0-1.
- Subjective feedback queries associated with System, Grammatical Category & Sentence level have been framed

The methodology is not dependent on any particular system and hence, can be used to evaluate translations from any system and can be extended to any language pair.

Development of Annotated Speech corpora for Indian Language (Hindi & Marathi):.Speech corpora involves selection of text data, design of minimal set of phonetically rich sentences, phrases & words consisting of multi-form units of speech, database format design and introducing grammatical syntax and context information with annotation. The minimum but sufficient text content selection has been made. Phonetically rich sentences consist of words having C03VC03 type monosyllables in maximum numbers. These syllables occur at starting, middle or at end positions in the word or it may occur in isolation. A typical set of 500 phonetically rich sentences were extracted from GyanNidhi corpus using Vishleshika software. To generate high quality synthesized speech, a vocabulary of unique 250 words has been created to cover words related to day, month, year, time, Quantitative Units, currency other than 1000 most frequent words & 1000 most frequent clustered words etc.

The type of sentences influences the prosodic patterns. A set of about 1000 prosody rich sentences reflecting anger, joy, and sadness, question type sentences, negative, command, exclamations etc. is created with the help of linguists. Recording has been done by Professional speakers (Male & Female) environment in a noise free & echo cancelled studio, at a sampling rate of 44.1kHz (16 bit) in stereo mode. Speech units are tagged in a hierarchical manner at sentence, word, syllable and phoneme levels.

Gyan Nidhi: A multilingual parallel corpus has been developed for English & 11 Indian languages. Parallel text contained in books (translated version in more than one language) is stored in UNICODE format. For corpus management, an application has been developed to enable user to get information on no. of pages, Author, abstract, keywords information & no. of languages in which the book is parallel. It displays the aligned text for selected set of languages. The metadata of the files is stored in the form of XML. This corpus is useful in development of multilingual dictionaries, spell checkers, creating translation memory for EBMT Systems. Sources of text are Sahitya Akademi, SABDA Pondicherry ,Navjivan Publishing House, Publications Division, Pustak Mahal, NBT India etc.

Devanagari OCR: OCR for Devanagari script has been developed (based on technology of ISI Kolkata) and test data was taken from Digital Library. A training module was added to the OCR "Chitraksharika" which can be trained for recognizing character glyphs which are different for different fonts & complex characters slowly disappearing in current writing style. In collection of documents for Indian language content, printing technologies varied from typesetting to cyclostyled document, block printing and laser setting which means that OCR system software spell checker and dictionary support needs to be rich and diversified. These modules are also being added.

Digital Library: Digital Library Mega Centre project is aimed to digitize 14 Million Pages of rare books, manuscripts, magazines etc. for putting on the web for reading by people. The task is being carried out at various centres across the country namely ICCR, IARI, Nagari, Gurukul Kangri Vishvidyalaya, Hardwar, BITS Pilani, Association of Indian Universities etc. Other Digital library projects are being implemented at Nagari Pracharini Sabha Varanasi, Kumaon University Nainital, GB Pant University, Pantnagar .As the data in physical form is being digitized, the tools and utilities required for its optimal use are being worked that would help in managing, searching and maintaining the digitized information better.

Tools for Digital Library: Cross Lingual Information Retrieval, Text Summarization, Multimodal Interface for Digital Library Access, Searching and Indexing Tools

Projects mentioned above are being implemented at SNLP Lab,Centre for Development of Advanced Computing (C-DAC), Noida (a scientific society of Ministry of Communications & Information Technology, Govt of India). Most of the projects are funded by DIT, MoCIT.

Language Technology Research Laboratory, University of Colombo School of Computing

The Language Technology Research Laboratory (LTRL) at the University of Colombo School of Computing (UCSC) was established in March 2004 under the PAN Localization project funded by International Research and Development Centre (IDRC), Canada. The UCSC had been involved in both localization and local language computing, especially in Sinhala, through various government bodies and undergraduate and post graduated research. The areas where UCSC had been working on are namely, developing Sinhala fonts and keyboard drivers to display based on proprietary encodings, standardizing Sinhala for UNICODE, natural language processing tools.

Currently the LTRL is involved in developing some key natural language processing (NLP) resources to enhance its research on Sinhala language processing and NLP tools that are essential in human computer interaction and localization. Sinhala corpus of 10 million words and a Sinhala Lexicon of 30,000 words are significant among the resources that are being developed. Research and development is being done on Optical Character Recognition (OCR) system for Sinhala scripts and a Text To Speech (TTS) engine for Sinhala.

Sinhala Corpus

This corpus will eventually consist of 10 million words that are drawn from various genres such as Newspaper, Fiction, Academic & Scientific writing, religious writing, etc. The final corpus will be in UNICODE text and each file will have a header that provides relevant information about the text. Currently we have released a 600,000 word version of the corpus and have already collected more than 5 million words of text overall.

Sinhala Lexicon

This lexicon will consist of 30,000 words with morpho-syntactic and semantic information. Most of the lexemes will be given with their respective Tamil and English translation equivalents. It is hoped that this lexicon will be used both in NLP task and by ordinary users to as an electronic dictionary. Several existing resources such as electronic versions of the dictionaries in circulation have been exploited to enrich the content of this lexicon.

Text To Speech Engine

This component aims to develop a commercial grade TTS engine using di-phone concatenation. Currently a Syllabification Algorithm has been developed for Sinhala with very high accuracy. In addition a Letter to Sound module has also been completed.

Optical Character Recognition

Various image processing techniques are being developed and tested in this component in order to determine the most suitable technique for quick identification of Sinhala printed characters. While results obtained so far are very positive, scalability of the algorithms adopted is currently being tested. The final goal is to produce a commercial grade OCR system for printed Sinhala characters.

Other work

LTRL also recently undertook a project for the Information & Communication Technology Agency (ICTA) to define an official Sinhala Language Collation. Other work in which the LTRL is currently involved in includes playing an active role in defining an official Sinhala Language glossary for Computer Terms and the translation of the graphical user interface.

In addition to the above mentioned activities the LTRL assists various public and private organization to improve their language technology related activities by conducting workshops and training programmes on UNICODE, especially on website development using UNICODE, UNICODE font development, and database management using UNICODE data.

All research activities carried out by the LTRL are supported by a group of eminent scholars drawn from different linguistics traditions and various universities of Sri Lanka.

Report

Center for Research in Urdu Language Processing (CRULP)

Sarmad Hussain

Urdu is a rich language with a multilingual and multi-cultural heritage. Its roots in Arabic, English, Persian, Sanskrit and other languages give Urdu a diverse body of sounds and underlying linguistic structure. Similarly, multi cultural background introduces a varied tradition of calligraphy, prose poetry, and other forms of art in Urdu. This rich heritage makes Urdu far more computationally interesting and challenging than many other languages. This is also true for Pakistan's other regional languages including Balochi, Brahvi, Pashto, Punjabi, Sindhi, Siraiki, etc. These challenges pose a vast unexplored training field for researchers in computer science.

CRULP (www.crulp.org) was formed in 2001 at National University of Computer and Emerging Sciences (www.nu.edu.pk) with the following objectives:

- conduct research in the linguistic aspects of Urdu and regional languages
- participate in standardization efforts in Urdu and regional languages
- evolve computational models of Urdu and regional languages
- promote and assist in content development for Urdu and regional languages

To achieve these objectives, CRULP is actively involved in linguistic inquiry, standardization for language computing, script processing, speech processing and computational linguistics for these languages.

Due to unavailability of trained people in these areas in Pakistan, CRULP started offering the only under/graduate level program in Script, Speech and Language Processing in Pakistan. CRULP is offering specialized courses and trainings in this area. In addition, the center also has specialized software, labs, library and recording room for its researchers.

Research at CRULP is based on solid linguistic analysis. Most of the linguistic work being currently pursued is applied in nature, for eventual computational modeling, and most of it is focused on Urdu. Work is being done on fundamental issues related to character sets, diction and pronunciation conventions. This research has led to developing national standards for character-set and collation in Urdu and Sindhi and enhancing Unicode standard for Urdu. Work is also in progress in other languages. Additional work is also being done in phonetics (especially acoustic phonetics), phonology, morphology, syntax and grammar, and very limited semantics of Urdu.

Urdu is written in Nastaleeq style of Arabic script. This script is very difficult to model and recognize. CRULP is actively involved in developing intelligent fonts (e.g. Open Type Fonts) to realize Nastaleeq. Computational techniques are also being developed for line/word-breaking and for justification of text (which requires stretching and overlapping of text, instead of simple space insertion between words). In addition, research is also being actively done on developing new techniques for optical character recognition of printed text and handwriting recognition. Existing techniques based on Hidden Markov Models are being enhanced for this purpose. Finally, models are also being created for automated diacritization of Urdu text.

In addition to the work on script, there is research being done in modeling Urdu speech for development of speech synthesis and recognition systems. This includes, for example, letter-to-sound rules, statistical durational and intonational modeling, phonological syllabification and stress assignment of words and sentences, etc.

Finally, much work is being done in Computational Linguistics. The core of the work is to develop a computational lexicon of Urdu for other applications. The lexicon includes up to twenty dimensions for each word, including pronunciations, inflections, senses, English translation (at sense level), sub-categorization frames (for verbs), collocations, synonyms, idioms, etc. Under a project funded by Government of Pakistan, a lexicon of 12000 words has already been developed. Work is also being done in spell-checking (with rule-based models already developed and giving 90%+ accuracy), grammar-grammar checking and machine translation. For machine translation, CRULP has researched and developed its own parsing, transfer and generation engines for Lexical Functional Grammar (LFG) formalism. It has also developed grammars for English and Urdu and also transfer grammar for English-to-Urdu translation. Transfer lexicon is also being developed for this purpose. Most of the work being done is applied, but some theoretical work on machine translation has also commenced, focusing on transfer mechanisms within LFG.

CRULP is also playing a regional role in the development of language processing as a discipline in Pakistan and Asia (especially in developing countries) and for the development of local language computing (see www.PANL10n.net). CRULP offers MS and PhD programs in these areas and offers specialized coursework. The center has five faculty members and about 40 full-time researchers funded by national and international organizations.

Editor's Note

Welcome to Phuket!

On behalf of the Editorial Committee of the AAMT Journal, it is my pleasure to present to you this special issue of AAMT Journal for the Tenth Machine Translation Summit, MT Summit X. Because the MT Summit X is held in the South East Asia, we would like to use this opportunity to introduce to you research activities on machine translation in South and South East Asia, which may not be familiar to the participants from other regions.

I would like to thank everyone who submitted information of their regions to this issue. I would also like to express my gratitude to the members of the Editorial Committee for their great efforts. And I would like to thank Ms. Kayoko Takada, secretary of AAMT, and Mr. Yoshiaki Murakami of NAVIX, who did all the tedious but crucial work for this issue.

This issue consists of papers from various research groups in the region on their research activities. We hope you will greatly benefit from this special issue of AAMT Journal. We would be more than happy if you could find candidates for your collaborative research, and a direction for your future research on machine translation.

Hitoshi Isahara
Chair of Editorial Committee of AAMT Journal

AAMT Asia-Pacific Association for Machine Translation
Mitsui-Sumitomo Kaijo Bldg., Annex 3F
3-11, Kanda-Surugadai, Chiyoda-ku, Tokyo 101-0062
TEL.: +81-3-3518-6418 FAX: +81-3-3518-6472 E-mail: aamt@aamt.info

Application Form For Regular Member

Date:

Please complete the following blanks.

(Please print) Last Name First Name Middle Name

Name: _____

Mailing Address: _____

Phone (Home): _____ (Office): _____

E-mail: _____

Company Name and Address:

_____

Business Title: _____ Type of Business: _____

Signature: _____

Payment: Application fee: ¥1,000 (First year only)
Membership fee: ¥5,000 (Period: April 1–March 31)
Total: ¥6,000

How to pay:

＊Bank transfer: Mizuho Bank - Ochanomizu Branch, Tokyo, Japan
Account Name: Asia-Pacific Association for Machine Translation (AAMT)
Account Number: 1737479 (Ordinary A/C)

＊Credit Card: □VISA □MasterCard Expiry Date:
Card No.:
Card Holder's Name:

Note: Application is accepted with payment.
No refund or credit if you cancel after payment.
Applicants are required to pay all handling charges.

MT Summit X

The 10th Machine Translation Summit

September 12-16, 2005 : Phuket, Thailand

AAMT Journal
Special Issue
September 2005

Asia-Pacific **A**ssociation for **M**achine **T**ranslation (AAMT)
Mitsui Sumitomo Kaijo Bldg., Annex 3F
3-11, Kanda-Surugadai, Chiyoda-ku
Tokyo 101-0062 JAPAN
TEL : +81-3-3518-6418 FAX : +81-3-3518-6472
URL : http://aamt.info Email : aamt@aamt.info

Typeset and printed in Japan by :
Navix Co., Ltd.

Published in Japan by :
Asia-Pacific **A**ssociation for **M**achine **T**ranslation (AAMT)