5.4 Language Engineering (LE)

a) Grammar

5.4.1 Paninian Grammar Framework Applied To English

Bharati, Akshar, Vineet Chaitanya and Rajeev Sangal, *South Asian Language Review, Creative Books, New Delhi, 1997a.*

Abstract

Computational Paninian Grammar framework (PG) has been successfully applied to modern Indian languages earlier, using which anusaaraka machine translation system has been built (Narayana, 1994). In this paper, we show that PG can also be applied to English resulting in an elegant computational grammar. First, we generalize the notion of vibhakti to include position of the word in a sentence along with its case and associated preposition, if any. This allows us to use the familiar PG notions of karaka chart, karaka chart transformation, and sharing rules (Bharati et al., 1995) to account for the English actives and passives, lexical control, infinitives, etc. A transformation of the karaka chart and the vibhakti therein, very naturally accounts for what is called movement.

Second, we introduce a new vibhakti called TOPIC position (which corresponds to the first position in a clause) and a new operation called join for connecting a relative clause to its head. These two together handle long distance dependency in relative clauses and wh-questions, raising, tough-movement, pied-piping, etc. The karakas with TOPIC vibhakti appear at the beginning of the clause. This paper establishes that PG is more general than hitherto considered, and can be used to explain not just free word order languages but also positional languages. Further research is continuing on this and related aspects.

5.4.2 Examining The Syntactic Alternations Of Hindi Verbs With Reference To Morphological Paradigm

Debasri Chakrabarthi, Pushpak Bhattacharya, *Language Engineering Conference, University of Hyderabad, India, Dec. 2002.*

Abstract

The aim of this paper is to show the alternative pattern of verbs in Hindi. The work is guided by Beth Levin's work on English verb classes and alternations where English verbs are classified semantically according to their argument structure. There is a strong belief that the semantic nature of verbs is largely dependent on its argument structure. The nature of Hindi verbs shows that along with the argument structure, attention should also be paid to the phonological changes which influence the morphological paradigms. Preliminary attempts have been made to class the simple verbs of Hindi into different morphological paradigms along with the phonological changes. At present, the work is limited only to the syntactic structure. Semantic characteristics have not been dealt with. Focus is mainly given on nominal, transitive-causative, and intransitive-causative variants.

5.4.3 A Hybrid Approach To Pre-Conjunct Identification

Sebastian van Delden, *Language Engineering Conference, University of Hyderabad, India, Dec. 2002.*

Abstract

An algorithm that identifies pre-conjuncts in natural language sentences is described. Syntactic and semantic information is combined to ascertain the size of the syntactic relation that precedes a coordinate conjunction. The approach is not domain specific and relies only on part-of-speech information. Preference results on several corpora are given.

5.4.4 A Finite State Automation For The Oriya Negative Verbal Forms

Kalyanamalini Sahoo, *Language Engineering Conference, University of Hyderabad, India, Dec. 2002.*

Abstract

This paper discusses the processing of negative verbal forms in Oriya in a deterministic Finite State Automaton. A morphologically agglutinative language like Oriya has 'phrasal' or 'constituent' negation, where tense, aspect etc. impose restrictions on NEG marking. Negation can be marked by various NEG morphemes in various positions of the verbal form, but is marked only once. That is, the occurrence of a NEG morpheme restricts the occurrence of any other NEG marker in the verbal form. Such multiple positional slots for the NEG morpheme with respect to tense, aspect poses constraint for the processing of the string by FSA. The FSA being a unidirectional machine, cannot backtrack, and thus, cannot account for such mutual exclusiveness of the items if all the three NEG items are available in a single chart. So, to account for this problem, we propose different types of processing for the different positional slots of NEG morphemes.

b) Language & Script Analysis

5.4.5 Building A 'Free' Infra-Structure For South-Asian Languages

Bharati, Akshar, Amba P Kulkarni, Vineet Chaitanya, Rajeev Sangal, and G Uma Maheshwar Rao, *Proc. Of SAARC conference on Multi-lingual and Multi-media Information Technology, CDAC, Pune, 1-4 Sept. 1998a, (Keynote lecture).*

Abstract

It does not need to be stressed that it is important to provide knowledge and information in electronic form in South Asian (SA) languages. This task requires the development of software for searching texts, script conversion, dictionaries, spelling checkers, multi-lingual access software, etc., and of course, a rich collection of texts in electronic form. All this can be called the infra-structure for language.

There are a number of problems which need to be addressed.

(1) Very few word processors are following any standards regarding coding schemes while entering texts in SA languages. This renders the texts unusable across platforms. Even if another user has the right platform, the only thing he can usually do is to view the text. Normally he cannot even annotate the text using the keyboard. While the long term solution is for everybody to follow the ACII standard; in the short term, there is a need to develop code converters rapidly. This task has been automated to a large extent for Devanagari. The same should be done for other scripts.

(2) The technical feasibility of multi-lingual access software has been demonstrated. (Though the machine translation technology is far away.) Anusaaraka systems for accessing texts in five SA languages are under development, and alpha-version for some have been released. This task can be taken up at a wider scale covering all SA languages. The systems already built can also be refined further.

(3) Electronic texts and resources such as dictionaries, thesauri, lexical databases, are urgently needed. These can be prepared by the collective effort of myriads of people.

In this paper, we argue that the SA language infra-structure can best be developed through a large cooperative effort. The GPL “free” software model is best suited for this development because the source code is open, and license is given to all to refine and redistribute it. All the anusaaraka systems (developed under funding from DOE) are available as GPL free software with source code, for everyone to use and contribute to. As another cooperative effort, a free version of an online English to Hindi dictionary is expected to be available shortly.

5.4.6 Vishleshika: Statistical Text Analyzer For Hindi And Other Indian Languages

Sunita Arora, Karunesh Kr. Arora, S.S.Agarwal*, *Workshop on Spoken Language Processing, TIFR, Mumbai, January 9-11, 2003 .*

Abstract

The vast majority of knowledge and information is available in Natural Language and stored in the form of text in books, articles, reports etc. This Knowledge source needs to be converted into digital knowledge base for making it easily accessible through computers and networks and for using in development of Human Machine Communication Systems. Statistical Analysis of text can provide information about phonetic and linguistic description and structure of a given language which can be used for developing Knowledge based Language/Speech Systems for communication.

This paper describes the development of a software tool named Vishleshika for conducting detailed Statistical Analysis of Hindi language and adaptable to other Indian languages. Several types of Statistical Analysis from simple frequency counts and linguistic features to syntactic and semantic analysis could be done with the help of this package. The objective is to shift the burden of many linguistic decisions to the Statistical Analysis.

5.4.7 Bangla Script: A Structural Study

Chaudhuri, B.B. and N.S. Dash *(1998) Linguistics Today, 1(2): 1-28.*

Abstract

In this paper we have tried to analyze the shapes of graphemes used in the Bangla script (as noted in printed documents). The study has focused on the formation of graphemes, their structural changes in case of compound grapheme formation, contextual use of graphemes and allographs, statistical analyses of their occurrences, and their positional and functional roles in case of semantic change. The purpose of this study is to understand the role of graphemes in the language, to show their behavioral peculiarities, and if possible, to find out the reasons of such peculiarities. Information obtained from this study may by useful for Optical Character Recognition, Spell-checker designing, Key-board Designing, Cryptography, Language Teaching and Natural Language Processing in Bangla.

5.4.8 A Corpus-Based Study Of The Bangla Language

Dash, N.S. and B.B. Chaudhuri *(2001) Indian Journal of Linguistics. 20: 19-40.*

Abstract

The recent studies on natural language have shifted their attention from rationalistic approach to empirical approach, where people are more concerned in applying corpus-generated data-base in language

study than applying their intuitive assumptions. This change of attitude adds a new dimension to language study hitherto unknown to the linguistic world. Keeping this scenario at background, we examine how corpus can be used for studying Bangla language, and how new information and data obtained from corpus can be used in natural language processing (NLP) in Bangla. This study is entirely based on the findings from a sample monitor corpus systematically developed from Bangla data collected from various written documents published within a fixed time span.]

5.4.9 Bangla Pronouns - A Corpus Based Study

Dash, N.S. *(2000) Literary and Linguistic Computing, 15(4): 433-444.*

Abstract

Bangla is the second most widely-spoken language in the Indian subcontinent, yet has not been the focus of much research activity in either corpus linguistics or language engineering to date. This paper describes the automatic processing of pronouns in three and a half million words of Bangla corpus data. A corpus based analysis of Bangla pronouns is developed, and a new approach to the analysis of Bangla pronouns is taken as a consequence. On the basis of this analysis a system is then developed to identify and analyze Bangla pronouns in corpus data]

5.4.10 Corpus-Based Empirical Analysis Of Form, Function And Frequency Of Characters Used In Bangla

Dash, N.S. and B.B. Chaudhuri *(2001), Published in Rayson,* P., Wilson, A., McEnery, T., Hardie, A., and Khoja, S., *(eds.) Special issue of the Proceedings of the Corpus Linguistics 2001 Conference, Lancaster: Lancaster University Press. UK. 13: 144-157. 2001.*

Abstract

We explore various formal and functional aspects of Bangla characters used in a written text corpus designed systematically from data collected from various text documents published within 1980 and 1995. Study of characters is important to understand their structure and role, trace their behavioural peculiarities, and to know the reasons of such peculiarities. Here, we focus on formation of characters, their structural change in case of compound and cluster formation, their contextual use, their statistical occurrence, and their functional role in words. We also encompass role of punctuation marks used in Bangla texts. The corpus performs the role of an empirical data-base for necessary information. Shape of characters are analysed by classifying them into various groups to distinguish among graphemes, allographs, graphic variants, compound, and clusters. Characters in isolation as well as characters in string are separately considered because characters used within words often differ from their features noted in isolation. Statistics allows us to get a picture of frequency and rarity of particular character to determine its relative normality or abnormality in the language. Such study is useful in natural language processing, computational linguistics, optical character recognition, cryptography, key-board design, word-sense disambiguation, part-of-speech tagging, telegraphic code design, spell-checker design, dictionary preparation, machine translation, language teaching, etc.]

5.4.11 Using Text Corpora For Understanding Polysemy In Bangla

Niladri Sekhar Dash, Bidyut Baran Chaudhuri, *Language Engineering Conference, University of Hyderabad, India, Dec. 2002.*

Abstract

Polysemy implies presence of more than one sense of a particular word both in its context bound and context-free situation. The inherent aspect of polysemy is that a particular word will show multiple sense variations related by way of semantic extension and conceptual expansion. In last fifty years or so, polysemy is recognized as one of the central issues in lexical semantics, word sense disambiguation (WSD), actual sense extraction (ASE), language learning, conceptual categorization of words as well as in computer processing of language. Language users can identify a polysemous word quite easily, but are not equipped enough to decipher all its possible sense variations without appropriate reference to proper knowledge-base and other relevant information embedded with in contextual environments. Here we make an empirical effort to understand the basic nature of polysemy in Bangla. We also intend to know how words denote sense variations, which factors are instrumental in making them polysemous, what impact do they create in language understanding, and how sense variation can be best understood using information form various sources of knowledge base. Finally, we use a method to understand role of various contexts, obtained from Corpora, in maintaining an interface between words and their sense variations. With reference to Local Context is handy at certain times, but reference to Focal, Topical and Global contexts as well as to extra linguistic knowledge base is necessary for understanding sense variation and for obtaining actual contextual sense.

5.4.12 Natural Language Processing With Neural Networks

Qing Ma, *Language Engineering Conference, University of Hyderabad, India, Dec. 2002.*

Abstract

With learning-based natural language processing (NLP) becoming the main-stream of NLP research, Neural Networks (NNs), which are powerful parallel distributed learning/processing machines, should attract more attention from both NN and NLP researchers and can play more important roles in many areas of NLP. This paper tries to reveal the true power of NNs for NLP applications as supervised or unsupervised learning devices by concretely introducing three practical applications: part of speech (POS) tagging, error detection in annotated corpora, and self-organization of semantic maps.

5.4.13 Work On Indian Languages : A Perspective

P.V.S. Rao, *Proceedings of an International Conf.'Akshara-94' on Information Technology Applications in Language, Script and Speech, pp. 1-6, February 25-26, 1994.*

Abstract

Quite apart from the pros and cons of a major shift in India from use of English to that of Indian languages, it is clear that use of Indian languages for computer related activities poses a significant intellectual challenge. The Computer Systems and Communications Group of TIFR is involved in a number of application area concerning the use of Indian languages in the Computers and vice-versa. In several of these areas, significant progress has been achieved. This paper briefly summaries the current status of this work.

5.4.14 Multilinguality And Global Digital Divide

R.M.K. Sinha, *Joint IAMCR/ICA International symposium on the Digital Divide, November 16-17,2001, Austin, USA.*

Abstract

The new Information and Communication Technology (ICT) has given rise to a phenomenal growth in global electronic commerce, improved quality of life, health care, emergency interventions, international understanding, and is ushering in a knowledge-based society with more conscious, humane and better informed citizens. At the same time, the same technology is also responsible for pulling the sections of the societies apart in terms of "haves" and "haves-not", creating risks of conflicts, endangering peace and harmony. The economic conditions, getting translated into affordability of ICT in real terms, in most of the countries are largely responsible for this digital divide. However, an equally important factor creating the digital and knowledge divide, is the linguistic barrier. More than 85% of the Internet content happens to be in English whereas less than 10% of the world's population speaks English. Some of the major consequences of this linguistic barrier are: a real danger of death of the regional language (with the death of a language, its culture and cultural heritage also dies); imposition of English speaking culture in an unintentional manner; a bootstrapping effect in widening of the gulf between "haves" and "haves-not", due to a differential change brought in by the ICT in the economic condition and quality of life. The paper presents a deliberation on some of these issues.

5.4.15 The Dynamics Of Language Understanding

Kanagaluru Chandra Sekharaiah, D.Janaki Ram, *Language Engineering Conference, University of Hyderabad, India, Dec. 2002.*

Abstract

This paper presents a seminal, ongoing work on notion of linguistic schizophrenia and addresses the many consequent problems involved in the realm of machine intelligence. Emphasis is laid on research on idioms and phrases in natural language understanding.

5.4.16 Issues In Language Engineering In The Indian Context

Om Vikas, *Language Engineering Conference, University of Hyderabad, India, Dec. 2002.*

Abstract

Technological progress in knowledge exchange has occurred in three phases: from e-mailing in Arpanet (1965-85) to document browsing in the Internet (1985-2000) to concept navigation in the forthcoming Interspace (2000-2010). There is increase in returns in the knowledge economy which magnify the market leader's advantage. But the advancement in ICT is largely English centric resulting into sprawling digital divide. There is collapse in authorship, translation and quality of writings in other languages. UNESCO identifies challenges of multilingualism and universal access of information. Language engineering deals with construction and maintenance of tools for processing information in human languages. Translingual document retrieval is active area of research. Parallel corpora/translation memories with annotation are useful in machine translation. Machine translation technology has progressed since 1970s employing rule-based, example based, interlingua, transfer,

statical based and now focuses on natural language understanding. Speech technology similarly progressed since 1970s with focus on rule-based formant system to concatenated speech synthesis systems. Machine Translation and Speech Technologies are developing as complex system and hence there is trend towards cross project collaboration, synergy, critical mass, and deployable & scalable technologies. As technologies are moving from laboratories to market, systematic evaluation of these technologies is also being emphasized.

In India, development of technology for Indian languages may be categorized as A-Tech. Phase (1971-1990) with focus on technology adaptation; B-Tech. Phase (1991-2000) with focus on basic technologies when TDIL programme was launched; C-Tech. Phase (2001-2010) with focus on Creative and convergence Technologies. There had been achievements in terms of development of OCR, Text-to-Speech, Text editor, on-line machine translation system and content creation in most of the Indian languages. A number of basic tools are also available in public domain for wider use. Resource Centres have been set up for development of language technology based solutions. Some new initiatives are also being considered such as KUNDALINI (Knowledge, UNDerstanding & Acquisition of Language, INferencing and Interpretation), Speech-to-Speech translation and digital library initiatives. It is also proposed to initiate training programme at the master level in the computational linguistics to facilitate migration from linguistics and masters in the knowledge engineering to facilitate migration from engineering disciplines. Challenges ahead are multilingual open source software, localization and speech technology initiatives.

5.4.17 Annals Of Indian Language Computing

Om Vikas, *International Conference on Universal Knowledge & Language - 2002, Goa*

Abstract

India is a multi-lingual country with 18 constitutional languages and 10 different scripts. Eighteen constitutional Indian Languages are mentioned as follows with their scripts within parentheses: Hindi (*Devanagari*), Konkani (*Devanagari*), Marathi (*Devanagari*), Nepali (*Devanagari*), Sanskrit (*Devanagari*), Sindhi (*Devanagari/Urdu*), Kashmiri (*Devanagari/Urdu*); Assamese (*Assamese*), Manipuri (*Manipuri*), Bangla (*Bangali*), Oriya (*Oriya*), Gujarati (*Gujarati*), Punjabi (*Gurumukhi*), Telugu (*Telugu*), Kannada (*Kannada*), Tamil (*Tamil*), Malayalam (*Malayalam*) and Urdu (*Urdu*). These are in vogue in different states. There are less than 5 percent people who can work in English. Inspite of the plurality of languages and scripts, their script grammar and language grammars are quite similar, and they have 40 to 80 percent vocabularies in common.

Impact of Information Technology was felt as early in 1970s. Solutions towards adaptation of rapidly growing Information Technology for Indian languages were developed. Input-output problems and coding schemes were analysed. In 1990-91, Government of India launched the program on TDIL (Technology Development of Indian Languages) under which projects were supported for development of corpora, OCR, Text-to-Speech, machine translation and generic software for Information processing. Standards for keyboard layout and internal Code for Information Interchange were also evolved. Rapid changes in Information Technology (IT) - Operating systems, Generic packages, Peripherals, Internet and networking – made Indian solutions for IT adaptation drag behind. But demand by Government and people continued as thrust for developing Indian language technology solutions, especially, in the wake of establishing world–level par excellence by Indian Software Professionals and companies. In 2000-2001, Government launched mission-oriented program for Technology Development for Indian Languages (TDIL) with focus on seven major initiatives: Knowledge Resources, Knowledge Tools, Translation Support Systems, Human Machine Interface Systems, Localization, Standardisation and Human Resource Development for Language Technology. Thirteen Resource centres for Indian Language Technology Solutions (RC-ILTS) were supported covering all 18 Indian languages. COILNet Centers for content Development and IT Localisation are being set-up in Hindi speaking states to increase IT penetration and promote use of IT in sustainable socio-economic development. Indian Language Technology Vision 2010 has been prepared with the Vision statement " Digital Unite and Knowledge for All". Technology audit focuses on peer-review, peer-technology sharing and product–oriented technology development. India is voting member of UNICODE consortium. Industry consortium for Indian language technology has also been formed. Technology Business Meet was held in November 2001 where technology developers and prospective technology takers had dialogue. 43 Technology Handshakes were signed. Activities concerning e-Content creation, IT localisation, on-line gisting and summarization, OCR, Cross-Lingual Information Retrieval, on-line Machine Translation, are being promoted to ensure information access in cyberspace in Indian languages. OCR for major Indian languages is ready with character level accuracy above 97%, on-line English-to-Hindi Machine Translation System integrated with Text-to-Speech is also operational.