5.3 Knowledge Tools (KT)

a) Lexical Tool

5.3.1 Application Of Multilayer Perceptron Network For Tagging Parts-Of-Speech

Ahmed, S.Bapi Raju, Pammi V.S. Chandrasekhar, M.Krishna Prasad, *Language Engineering Conference, University of Hyderabad, India, Dec. 2002.*

Abstract

This paper presents a neural network based part-of speech tagger that learns to assign correct part-of-speech tags to the words in a sentence. A multi layer perceptron (MLP) network with three-layers is used. The MLP-tagger is trained with error back-propagation learning algorithm. The representation scheme for the input and output of the network is adopted from Ma et al. [6]. The tagger is trained on SUSANNE English tagged-corpus consisting of 156,622 words. The MLP-tagger is trained using 85% of the corpus. Based on the tag mappings learned, the MLP-tagger demonstrated an accuracy of 90.04% on test data that also included words unseen during the training. Results from our experiments suggest that the MLP-tagger combined with the representation scheme adopted here could be a better substitute for traditional tagging approaches. This method shows promise for addressing parts-of-speech tagging problem for Indian language text considering the fact that most of the Indian language corpora, especially tagged ones, are still considerably small in size.

5.3.2 Morphological Generator For Tamil

P. Anandan,Dr. Ranjani Parthasarathi & Dr. T.V. Geetha, *Tamil Inayam, Malaysia 2001.*

Abstract

Tamil is a relatively free word order language, the only constraint being that normally the verb comes at the end. This flexibility in word ordering is possible due to the morphologically rich nature of the language. Information such as case rules and auxiliary verbs indicating aspect, tense, mood are all conveyed through morphological attachments to the root, noun or verb. This makes morphological generation of Tamil words a challenging task. Another issue is that unlike English where number matching between noun and verb alone is necessary, in Tamil as in many other Indian languages person and gender matching between subject and verb is also necessary. A morphological generator designed for Tamil needs to tackle the different syntactic categories such as nouns, verbs, postpositions, adjectives, adverbs etc. separately, since the addition of morphological constituents to each of these syntactic categories depends on different types of information.

In this work a morphological generator has been designed for each of the syntactic categories and then combined to morphologically generate a complete sentence. The underlying morphological structure for a Tamil noun is as follows: plural suffix - if any, oblique suffix - if applicable, the euphonic suffix - if optional plus the case suffix. While generating the noun derivatives from the roots linguistic rules determining the form of a plural suffix has to be considered. The attachment of case suffixes to nouns is an important part of the morphological generator for nouns. In addition, this has to take into consideration the fact that certain nouns can take case suffixes only in oblique form. The euphonic suffix sometimes comes along with oblique suffixes or with plural suffixes. This has also to be considered. During the combination of the root - noun with the above mentioned suffixes, “sandhi rules” have to be taken in to account.

The Morphological structure of Tamil verb is quite complex since it caters to person, gender, and number markings and also combines with auxiliaries that indicate aspect, mood, causation, attitude etc. While morphologically generating the verb, the gender, number and person of the subject is necessary in order to select the appropriate suffix catering to the selected tense. The linguistic rules determining the combination of auxiliaries with the verb is also quite challenging since more than one auxiliary can attach itself to the verb as suffixes. The combination becomes exponentially large, hence we have used semantic based heuristics to prune the combinations. While using auxiliaries zit is the last auxiliary that matches with the person, number, gender of the subject and not the root verb of the combination. Similar linguistic rules are used to generate derivatives for other syntactic categories also. Thus a Tamil morphological generator for different syntactic categories has been designed and implemented. This morphological generator can be used for *suggestion list generation of a spell checker* and as *a basis for English to Tamil translation*.

5.3.3 Computer Parsing Of Bangla Verbs

Chaudhuri, B.B., N.S. Dash and P.K. Kundu *(1997) Linguistics Today 1(1):64-86.*

Abstract

This paper deals with automatic parsing of Bangla verbs. It includes identification of verbs in texts, define their roots and suffixes, analyze their declensions, detect their meaning, and determine their role in sentence. Roots are divided into 27 sub-groups according to their structure while suffixes are sub-

grouped depending on their conjoining with particular root. Grammatical properties are encoded in 1 byte of binary string to tag with suffix. Entries in root lexicon are tagged with information about their form and correspondence with suffix. When a verb form is encountered, a stripping algorithm is invoked to find root-suffix pairs. Next, a Boolean mapping algorithm is used for valid concatenation (grammatical mapping) between two parts. If a valid parse is obtained the result is presented by decoding encoded information attached to root and suffix. For robustness and accuracy the process is run on a large collection of verbs compiled from a corpus proportionally chosen from published documents between 1981-1995. Information may be used for analysis of contexts in sentences.

5.3.4 A Gurmukhi Collation Algorithm

G S Lehal, *communicated to International Journal for Communication*

Abstract

Sorting and Indexing is one of the basic necessities of the database management system. But unfortunately there does not exist any software for automatic sorting of Gurmukhi words. The collating sequence provided by UNICODE or ISCII is not adequate, as it is not compatible with the traditional sorting for Gurmukhi words. In Gurmukhi unlike English, consonants and vowels have different priorities in sorting. Words are sorted by taking the consonant's order as the first consideration and then the associated vowel's order as the second consideration. In addition the properly sorting "characters" in Gurmukhi often requires treating multiple (two or three) code points as a single sorting element. Thus we cannot depend on character encoding order to get correct sorting instead we have to develop using sorting rules of Gurmukhi linguistic collation function which convert the word into some intermediate form for sorting. The Gurmukhi collation algorithm developed takes care of following two main factors into consideration:

1. Deciding the collating sequence.

2. Taking care of primary, secondary and tertiary weight levels to be assigned to some characters based on alphabetic ordering, semi-vowel ordering and ignorable character based ordering respectively.

The collating sequence for Gurmukhi characters has been developed after discussions with linguists and studying in detail the alphabetic order in Punjabi dictionaries. The collating sequence takes care of vowels, consonants and conjunct consonants. For assigning the secondary and tertiary weight levels, semi-vowels and ignorable characters such as hyphens have to be taken care of. Since we have multiple-level comparisons to be carried out, so first the text to be sorted is transformed into a collation element table and then into equivalent sort keys. These sort keys might consist of a string of base weights followed by strings for weights used for secondary and tertiary differences. These keys can then sorted based on some sorting algorithm.

5.3.5 Enhanced Version Of Morphological Analyzer And Parser For Tamil Language

U. Madhupriya, *MCA Project, Anna University, Chennai.*

Abstract

Natural Language Processing is one of the central domains of investigation in artificial intelligence. Crucial to the progress in this domain is a better understanding of the properties of natural languages and the development of linguistic formalisms both for analyzing these properties and for computer applications.

Morphology can be defined as an internal structure of words. A Morphological analyzer breaks a word into its root word and associated morphemes. A morpheme is defined as the smallest part of a language that can be regularly assigned a meaning.

Tamil is a morphologically rich language in which most of the morphemes coordinate with the root words in the form of suffixes. Person, gender and number markings of the subject of the sentence. In addition auxiliaries, which convey modal, aspect, etc. also combine with the main verb and form a cohesive unit. Unlike most other languages, in Tamil, case markers occur along with the nouns and number markings. Hence a tool for morphological analysis of Tamil language is necessary.

Most of the languages of the world allow considerably more variation in word order than does English. The subject and object are identified not by their positions but by their inflectional endings. Parsing of Tamil sentences is a requisite for various applications. The syntactic correctness of the construct can be known from parsing. In a parser, morphological analysis of words is an important prerequisite for syntactic analysis. Properties of a word the parser needs to know are its part-of-speech category and the morpho syntactic information encoded in the particular word form.

The morphological analyzer and parser are built using Jlex – a lexical analyzing tool and java Cup – a parsing tool.

5.3.6 Latent Semantic Analysis And Applications For Tamil Documents

G. Gowri Mani, *MCA Project, Anna University, Chennai.*

Abstract

Implementing latent semantic indexing and various applications using LSI for Tamil documents is the main goal of the project. Here, two different applications are considered for the purpose. They are Information Retrieval using LSA and Intelligent essay assessor.

Information retrieval is aimed at automatically retrieving information, based on the semantic similarity of the documents, rather than on the perfect word match. Unlike most of the information retrieval systems of present, this system is aimed at retrieving document, which are conceptually similar. The system is developed as a web-enabled product, enabling any user to readily download the product from the net and use it to the fullest. The system automatically, extracts the higher order associations from the given document. Further it converts them to a semantic space, mathematically retrieving the conceptual details of the document. This semantic space is used to retrieve the related document, based on a user query.

The Logical sub modules involved are,

- Concept computing of given documents
- Query processing and document retrieval
- Updation of Documents

The Intelligent Essay Assessor (IEA) is a software tool for scoring the quality of essay content. The IEA uses Latent Semantic Analysis, which is both a computational model of human knowledge representation and a method for extracting semantic similarity of words and passages from text. Simulations of psycholinguistic phenomena show that LSA reflects similarities of human meaning effectively. This system automates the concept gathering of any given corpus. It further is expected to evaluate the essay as consistent and efficient as that of a human. It applies various strategies to arrive at the final score of any given essay.

The logical sub modules involved are,

I. Concept computing of the given corpus

II. Scoring the essay based on various criterion involved.

5.3.7 A Grammar Tool For Sentence Generation And Cross-Language Communication

P.V.S Rao, *describing work done at the CSR Lab., Virtual conference WWDU 2002 – the 6th International Scientific Conference on Work With Display Units – May 23, 2002, at Berchtesgaden – Germany*

Abstract

This paper describes a grammar tool which can be a literacy aid as well as a facilitator of cross language communication at the national level (for multilingual countries) and at the global level. It enables a user to generate:

a) fairly complex sentences in a language that he is barely familiar with, and

b) equivalent sentences in a second language with which the user need not even be familiar with.

He can incrementally convey the intended 'concept' (underlying the sentence) to the machine in a non-sentential form. This is internally represented as a structure, which is not language-specific. This structure can be converted (using appropriate grammar rules) into sentences of a language.

The tool can be used in two modes:

a) as an aid for sentence generation in real life or

as a training tool for gaining competence, either in the first language of the user or in a second language.

5.3.8 Anaphora Resolution For Malayalam And Hindi

L.Sobha, *Unpublished Doctoral dissertation submitted to MG University, Kottayam.*

Abstract

It is well known that natural languages contain anaphoric expressions, gaps and elliptical constructions of various kinds and that understanding of natural languages involves assignment of interpretations to these elements. Therefore it is only to be expected that natural language understanding systems must have the necessary mechanism for the same. Most of the anaphora resolution systems developed so far seem, to the best of our knowledge, to be monolingual, i.e., the systems are language specific and it has not always been shown that these are easily extendable to deal with other languages. VASISTH (the anaphora resolution system), in contrast, is a multilingual system, which presently handles two languages from two different language families: Malayalam, from, Indo-Dravidian and Hindi from Indo-Aryan. It can easily be extended to handle other Indian languages as well, more generally, other morphologically rich languages. What further distinguishes VASISTH from other similar systems is that exploiting the morphological richness of the Indian languages, it makes limited use of grammatical rules and uses only morphological markings to identify subject, object, clause etc. It uses limited parsing:

the information required from the parser is limited to parts of speech tagging, clause identification, subject of the clauses and person-number-gender of the NPs. Initially VASISTH was developed and tested for Malayalam, and then modified for Hindi. It is well known that pronouns often have more than one possible antecedent: the pronoun resolution mechanism of this system captures the ambiguity but does not resolve it. The system aims to resolve (and achieves considerable success in doing so —- complete success as far as reflexives, reciprocals, and distributives — are concerned) the antecedent problem of referentially dependent elements such as pronouns — both anaphoric and cataphoric uses —, including the so-called "one – pronouns", reflexives, emphatic and non-emphatic, reciprocals, and distributives, gaps, and certain kinds of ellipsis. The parser and the anaphora resolution system work in "c".

5.3.9 VASISTH An Anaphora Resolution System For Indian Languages

Sobha,L and B.N.Patnaik *(2000) International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications ACIDCA'2000, Monastir, Tunisia.*

Abstract

The paper presents "vasisth", an anaphora and ellipsis resolution system, developed originally for Malayalam, an Indo-Dravidian language, and then tested for Hindi, an Indo-Aryan language. The testing was done to see if the system would apply without modification to another Indian language, structurally similar to Malayalam in many crucial respects, and if modifications are needed, what these modifications specifically are. The test gave encouraging results: only a very few minor modifications are needed for the system to apply equally efficiently to Hindi. One fact about the computational grammar that is implemented here which deserves attention is that it does not use more sophisticated notions of modern formal linguistics, and achieves its goal with very familiar concepts such as clause, subject, object, etc., which are identified with the help of morphological information, and concepts such as precede and follow. The more complex notion of hierarchy is not used. The result is quite encouraging: a very simple grammar, from the point of view of implementation, and the coverage is not affected. The algorithm developed here works on a partial parser, which is because the need for a complete parser has been eliminated for the operation of the system. The system works with a fairly high degree of success and it can be said with some confidence that it can be extended to other morphologically rich languages.

5.3.10 Vasisth An Ellipsis Resolution In Malayalam And Hindi

Sobha, L and B.N. Patnaik. *(2000), MT 2000 – Machine Translation and Multilingual Applications in the New Millennium, University of Exeter, United Kingdom: 19-22 November 2000*

Abstract

The paper presents an algorithm which resolves elliptical constructions for Malayalam, an Indo-Dravidian language, and then tested for Hindi, an Indo-Aryan language. The algorithm is a part of an anaphora resolution system called VASISTH. The testing was done to see if the system would apply without modification to another language, structurally similar to Malayalam in many crucial respects. The test yielded encouraging results. The computational grammar implemented here uses very familiar concepts such as clause, subject, object etc., which are identified with the help of morphological information, and concepts such as precede and follow. The algorithm works on partial parser.

5.3.11 Vasisth An One-Pronoun Resolution Algorithm For Indian Languages

Sobha, L and B. N. Patnaik. *(2000), Third International Conference on Discourse Anaphora and Anaphor Resolution (DAARC2000), Lancaster University, United Kingdom 16 – 18, November 2000*

Abstract

The paper presents an algorithm, which resolves one-pronoun construction for Malayalam, an Indo-Dravidian language, and then tests it on Hindi, an Indo-Aryan language. The algorithm is a part of an anaphora resolution system called VASISTH. The testing was done to see if the system would apply without modification to another language, structurally similar to Malayalam in many crucial respects. The test yielded encouraging results. The computational grammar implemented here uses very familiar concepts such as clause, subject, object etc., which are identified with the help of morphological information, and concepts such as precede and follow. The algorithm works on partial parser.

5.3.12 Object Oriented Design Of Semantically Driven Tag Based Hindi Sentence Analyzer

Om Vikas, *Conference on Computer Processing of Asian, Languages (CPAL-2, March 1992*

Abstract

The paper discusses the design methodology of a semantically driven TAG (Trea Adjoining Grammar) based sentence analyzer which analyzes simple and complex sentences, and performs syntactic and semantic checks having access to a Semantic Relational Lexicon to facilitate semantically driven parsing. Sentence is an expression of an event or a situation that an action (A) takes place on some entities (E) which may play a role of agent, object, instrument, etc. Five E:A relationships have been specified. Five syntactic classes have been included in this paper. EP and AP are termed as Entity Phrase and Action Phrase respectively. Class I: AP-Identification Procedures (+Modality), Class II: EP-Identification Procedures, Class III: E-Inflexion Procedures, Class IV: A-Inflexion Procedures, Class V: E-A Agreement Procedures. Tree Adjoining Grammar (TAG) theory may be used for combining sentences and clauses. EP may consist of EP and the J_Phrase (starting with 'jo', 'jisne', etc). To form ('kyaa') question, K_phrase may be added in the beginning or K_phrase (starting with 'kyaa', 'kisko', 'kaise', etc.) may replace EP/AP. Two sentences may be linked to an appropriate Epi of the main sentence through L_par that is linking parsarg (e.g. 'ki', 'kyonki', etc…). The proposed analyzer analyzes a sentence into entities (E), actions (A) and their modifiers with their features. The case relations are assigned from left to right in a sentence by using six rules of identifying Agent, Object, Instrument, Goal, Source and Locus (time, situation, manner).

b) Utilities

5.3.13 Bridging The Gap Between Home Dialect And School Language Through Multimedia

N. Anbarasan, *National seminar on Research and Innovations on Home and school language issues.*

Abstract

Language learning is a complex and time-consuming process. When it comes to Language learning pupil experience the difference between Home dialect and the school language at various linguistic levels such as phonological, grammatical, lexical and semantic. The dialect is acquired at out and out through the socialization process, where as school language is learnt in a formal schooling situation.

A homogeneous classroom is not guaranteed as pupils coming from different corners having diversity in language, culture, socio economic, tradition, custom, caste, religion etc. With this background, pupils respond and re-act differently to the same situation. Multimedia helps in preparing an unambiguous material. As different individuals require different types of teaching and learning suitable to their individualism, multimedia has the scope for configuring to meet these requirements. This would ensure better results and expected outcome. Multimedia helps one to get conformed with the pace of progress, one maintains and to accelerate the rate of achievement and ensure proper attainment.

In any formal language learning/teaching situation, there is a necessity to have supplementary materials with a view to :

(1) Re-inforcement of language learning

(2) Coverage of certain aspects which have not otherwise covered in the conventional material

(3) Supply additional information/teaching inputs

(4) Overcome certain deficiencies in teaching as well as materials.

The multimedia based Teaching/Learning materials (Aids) bring life and helps pupils better understand the abstract difficult concepts easier and grasp. It also motivates the learning interest of the pupils with the real life animations with sound effects and music in it and helps reduce the labour of the teachers. Multimedia also enables the learners in re-inforcement of the subject by means of games and in turn paves ways for debates and discussions.

5.3.14 Constraints In Developing Language Software

N. Anbarasan, *Seminar International Seminar on Tamil Computing, Chennai.*

Abstract

When the computers are penetrating deep into every field of science and society, it quietly shows up its English face and puts the vernacular user to a great disadvantageous position and are left out. It happens due to the limitations of the existing operating systems, tools and application software developed.

Any software is meant for processing of the given data from some source and produce some results. It also applies to any language software. The three components of a program namely Input, Processing and Output, normally developed in a complicated way due to the complicated standards apart from the complexities of the languages themselves.

Developing software to take care of inputting Indian languages is a complex process in the requirements to develop it for different kinds of users namely Typist, Non-Typist, Casual and Professional users, all of them need different types of input methods.