a) Corpora

5.2.1 Some Observations On Corpora Of Some Indian Languages

Bharati, Akshar, Sushama Bendre, and Rajeev Sangal, *Knowledge-Based Computer Systems, Tata McGraw-Hill, Dec. 1998c.*

Abstract

This paper presents some data from seven Indian languages, as extracted from machine readable corpora. In particular, two kinds of data are extracted: (1) Number of distinct high-frequency words in a language needed to achieve a designated level of coverage of a random text. (2) Percentage of common words occurring in different languages. (A word is defined as a sequence of alphabetic characters delimited by space and punctuation marks.) The first data follows the expected trend that south Indian languages have much larger number of distinct words because they are more inflected than the north Indian languages and agglutinative in nature. Telugu turns out to be more so than Kannada. The second data shows that languages in adjacent regions have many more common words, as expected. Aim of the paper is to introduce scholars to the possibilities of using computers for extracting useful data about languages from machine readable corpora. Data presented here has been used to build a software for detecting the language of an unknown text.

5.2.2 Corpus Generation And Text Processing

Dash, N.S. and B.B. Chaudhuri *(2002) International Journal of Dravidian Linguistics. 31(1): 25-44.*

Abstract

The introduction of corpus linguistics raises a strong defiance against the rationalists who argue that empirical language study is a skewed approach which deserves little admiration. However, advancement of computer technology, availability of huge language data-base, and application of computer in language research (in storing, sorting, processing, retrieving etc.) confirm sustainable growth and durability of corpus linguistics. Such resent developments open up many new and exiting areas of language study hitherto unknown to the linguistic community. The findings from corpora defy the observation of the rationalists by supplying new evidence for theory making as well as language research. Here we make an attempt to present an overall scenario of corpus linguistics with its birth and growth, issues related to its generation and dif ferent techniques applied for its processing.

5.2.3 The Process Of Designing A Multidisciplinary Monolingual Sample Corpus

Dash, N.S. and B.B. Chaudhuri *(2000) International Journal of Corpus Linguistics. 5(2): 179-197.*

Abstract

This paper discusses the approach of developing a sample of printed corpus in Bangla, one of the national languages of India and the only national language of Bangladesh. It is designed from the data collected from various published documents. The paper highlights different issues related to corpus generation, data-file preparation, language analysis and processing as well as application potentials to different areas of pure and applied linguistics. It also includes statistical studies on the corpus along with some interpretation of the results. The difficulties that one may face during corpus generation are also pointed out.

b) Dictionary

5.2.4 A Study Of Data Structures For Implementation Of Punjabi Dictionary

G S Lehal and Kulwinder Singh, *Cognitive Systems Reviews and Previews,* J. R. Isaac and K. Batra *(Editors), ICCS '99, Delhi, Phoenix Publishing House Pvt Ltd, pp. 489-497, (1999)*

Abstract

In this paper, the implementation issues involved in an electronic dictionary for Punjabi spell checker have been studied. The commonly used data structures for English dictionary have been modified to suit the non-linear nature of Punjabi. Five data structures have been implemented for a Punjabi dictionary of about 26,000 root words. These data structures are 1)binary search tree 2) trie 3) ternary search tree 4)multi-way tree and 5) reduced memory method tree. For experiment a text of 20,000 words was used, which consisted of 10,000 valid and 10,000 invalid words. It is observed that Binary Search Tree is the most suitable data structure in terms of memory consumed and time

taken for successful as well as unsuccessful search for words. The limitation of Binary Search Tree is that it is very difficult to offer a suggestion list or to look up all words differing by one or two characters. This limitation can be removed by using Trie structure. In case of trie, too, the trie in which each node contains the information of the Laga and Lagaakhar associated with each consonant is more memory efficient than the trie in which each consonant, Laga and Lagaakhar are separately stored at each node. The processing time for both tries was same for successful as well for unsuccessful searches. The ternary search tree was more memory efficient than the trie in which each node stored one character, but is inferior to that trie in terms of searching speed. The multi-way tree is the bulkiest and slowest data structure while reduced memory tree also does not provide much improvement, the reason being that in case of Punjabi the tree is 56 way as compared to 26 way tree in English and thus space is need and to store all 56 possible paths and consequently more time is taken to search and traverse the correct path.

c) Lexical Resources

5.2.5 LERIL : Collaborative Effort For Creating Lexical Resources, Proceedings Of Workshop On Language Resources In Asia

Bharati Akshar, Dipti M Sharma, Vineet Chaitanya, Amba P Kulkarni, Rajeev Sangal, Durgesh Rao, *along with NLPRS-2001, Tokyo, 27-30 November 2001*

Abstract

The paper reports on efforts taken to create lexical resources pertaining to Indian languages, using the collaborative model. The lexical resources being developed are: (1) Transfer lexicon and grammar from English to several Indian languages. (2) Dependency tree bank of annotated corpora for several Indian languages. The dependency trees are based on the Paninian model. (3) Bilingual dictionary of 'core meanings'.

5.2.6 Building Lexical Resources

Bharati, Akshar, and Dipti M Sarma, *Proc. IRIL-99: Information Revolution and Indian Languages, 12-14 Nov. 1999, Hyderabad.*

Abstract

Lexical resources in electronic form are extremely valuable and need to be prepared for all Indian languages. They are needed to develop applications such as machine translation, and many others. However, the development of such resources has to be done carefully, keeping the application in focus, otherwise the effort is likely to be wasted.

For example, while building lexical resources to be used by a machine translation system from Telugu to Hindi, the application must be kept in focus. It is best to do it using the bilingual approach. In this approach, senses of a word can be identified based on differences between the two languages, rather than in the abstract. Working out such differences is thus grounded on hard data. And then the differences that ought to be specially tackled are where the machine has difficulty. This also helps in keeping the work focussed.

One should also note an important aspect of the electronic resources is that there is nothing like press ready copy. Since the electronic data can be refined and updated and then distributed very easily and also it is very easy to record the changes made, one need not wait for the 'finished' entry. Availability of such a resource in electronic media also makes it easy for others to participate in enhancing or improving it.

In this paper, we take some examples of building lexical resources in the context of English-Hindi anusaaraka. We present here an example to illustrate different aspects in representing the knowledge about lexical items.

Examples for 'up'

1. E: run up the stairs, (source: OALD)
   H: (सीढी से) ऊपर जाओ,
2. E: look up in the sky
   H: (ऊपर) आसमान में देखो
3. E: go up in the building
   H: बिल्डिंग में ऊपर जाओ

In all these sentences substitution of 'up' by 'ऊपर' in Hindi seems to yield good sentences in Hindi. Let us look at the word by word substitution in these sentences

दौड़ो ऊपर ^सीढी (के)

देखो ऊपर में ^ आसमान

जाओ ऊपर में ^ इमारत

But the following example shows that only 'ऊपर' cannot capture the full meaning from English. Instead 'के_ऊपर' captures it better. Therefore, we add an optional 'के' before 'ऊपर' to substitute 'up'

3. E:she tried her best to climb up the tree

H: उसने पेड़ के_ऊपर चढ़ने का भरसक प्रयास किया

Other possibilities in Hindi

H : उसने पेड़ पर चढ़ने का भरसक प्रयास किया

Now look at the following example:

4. E:we ran up the hill (source: Q&B)

H: हम पहाड़ पर ऊपर की ओर दौड़े

Here we find that there are instances in English where 'up' has a sense of 'movement towards, i.e. upwards'. 'के_ऊपर' by itself would fail to incorporate this sense. Therefore, it is suggested that we extend 'के_ऊपर' as' > [के]_ऊपर [की _ओर] Anusaaraka output for these sentences would be:

1.@ : दौड़ो > [के]_ ऊपर _[की _ओर]^^^सीढ़ी{ब.}

: दौड़ो सीढ़ी {ब.} [के]_ऊपर_[की_ओर]

2.@ : जाओ > [के]_ ऊपर _[की _ओर]>में^^^सीढ़ी{ब.}

3.@ : वह{स्त्री} कोशिश_किया उसका {स्त्री.} सबसे_अच्छा-–चढ़ना

> [के]_ ऊपर _[की _ओर]^^^पेड़

4.@ : हम दौड़े > [के]_ ऊपर _[की _ओर]^^^ पहाड़

However, when we come across sentences such as the following,

our formula appears to fail.

5. E: further up the valley (source: OALD)

H: घाटी में और आगे

Here the meaning is not 'upwards' in the valley but 'ahead' in the valley. Similarly :

6. E: walk up the road (source: OALD)

H: सड़क पर आगे जाओ

Therefore, we add आगे in our Hindi equivalent of 'up'.

>[के]_ ऊपर _[की _ओर]/ आगे

7. E: sail up the river

H: नदी की धारा के विरूध नाव चलाओ

H: नदी में ऊपर की ओर नाव चलाओ

@H:नाव{पाल_की}चलाओ_के_ऊपर_[की_ओर]नदी

Final solution for the English preposition 'up'

>[के]_ ऊपर _[की _ओर]/ आगे

Issues in the design of anusaaraka dictionary and use of appropriate notation have been discussed elsewhere. Even if we were to ignore those, the example above illustrates the nature of detailed work that needs to be done in building lexical resources. Most important lesson from our experience is that the building of the lexical resource should be tied to a concrete application at the start itself.

It has been argued elsewhere, that the initial part of this voluminous work should be done by involving thousands of non-experts. Ordinary bilingual people with an aptitude for language, can use a monolingual resource for English (such as OALD: Oxford Advanced Learners' Dictionary) to supply Hindi equivalents, or example Hindi sentences illustrating the uses of different senses mentioned in the mono-lingual resource.

5.2.7 The VOLEM Project: A Framework For The Construction Of Advanced Multilingual Lexicons

Ana Fernandez, Gloria Vazquez, *Language Engineering Conference, University of Hyderabad, India, Dec. 2002.*

Abstract

We report in this short document the results of a Regional European project carried out on Spanish, Catalan, Occitan and French whose aim is to design a lexical knowledge base where syntactic and semantic descriptions have been normalized and are treated in a uniform way cross-linguistically. Besides the scientific aspects, one of the aims is to make less developed languages such as Occitan or Catalan accessible on the WEB to a larger audience.