MEDIA LAB ASIA
Mumbai Hub

AGRO-EXPLORER

PROJECT REPORT

Under the guidance of :-

Prof. P. Bhattacharya
Deptt of Computer Science
and Engineering,
I.I.T, Bombay.

Project Interns :-

Sarvjeet Singh*
Tushar Chandra§
Upmanyu Misra§
Ushhan D. Gundevia§

* Deptt. of Comp. Sc. and Engg, I.I.T, Bombay.
§ Deptt. of Comp. Sc. and Engg, I.E.T, Kanpur.

1. Abstract:-

Agro-Explorer is a meaning based, inter-lingua search engine designed specifically for the agricultural domain. The techniques used are very generic and hence can be applied to any domain. In this report, we discuss its features and techniques. The need to develop a meaning based, inter-lingual search engine and its underlying principles are provided. A brief discussion of UNL, the underlying intermediate language, is also given. Furthermore, the strengths of our search methodology, with respect to other techniques, are highlighted. We also mention further enhancements that may be implemented to greatly enhance Agro-Explorer's usability for the masses.

2. Aim:-

To develop a search engine that would perform meaning-based search over a document database and return the relevant documents, irrespective of their language. Also, it should indicate the relevance of the searched results to the input query.

3. Theory:-

English, the chief communication language worldwide has forced many non-English speakers to spend time and resources over learning it. Despite this, the handicap still exits. Consequently, people have little chance to access other languages, such as, Spanish, Hindi, French which are otherwise rich and culturally distinct languages. Also, because of this, the power of Internet, as the best repository of information, is dampened. Whatever queries are searched on the Internet are in English and searched only on English documents, thus ignoring the vast information bank that is available in languages other than English.

India is primarily an agricultural economy, with eighty percent of its population still living in the villages. Majority of the farmers have low levels of literacy and very few avenues for authentic information. To eliminate this gap, they should be able to access the vast amounts of information available on the web. Moreover, almost none of them have any knowledge of English. Secondly, they don't have the patience and the time to search hundreds of pages to find relevant information. Hence, it is imperative that the search results returned are all highly relevant and in a language that the user understands.

3.1 Universal Networking Language [1]:-

To overcome these handicaps, we use the Universal Networking Language (UNL), an electronic language for computers to express and exchange every kind of information. The UNL represents the meaning of a sentence through a hyper graph having concepts as nodes and relations as arcs.

Binary relations are the building blocks for UNL documents. They are made up of two UW's and a relation

<Binary Relation>::= <Relation Label>| ":"<Compound UW-ID>| "("{<UW1>| ":"
<Compound UW-ID1>} ","{UW2>| ":"<Compound UW-ID2>} ")"

Where:

Relation Label	String of 3 lower case alphabets. Ex. **agt**, **mod**, **obj**, etc.
Compound UW-ID	String of 2 characters identifying each instance specified by the compound UW.
UW	Character strings representing concepts.

Universal Words are character strings representing concepts. They are annotated with attributes to that provide information about how the concept is being used in a particular sentence.

<UW>::=<Head Word>|<Constraint List>|| ":"<UW-ID>|| "."<Attribute List>

Where:

Head Word	An English word interpreted as a label for a set of all the concepts that correspond to that word in English.
Constraints	Restrictions on the interpretation of a UW to a specific concept.
Attributes	Provides information on how the concept is being used. Ex. **@past**, **@plural**.
UW-ID	Used to indicate some referential information.

There are four types of UW's

Basic UW	Character strings corresponding to English words. Ex. liquid, state.
Restricted UW's	Denotes a specific concept. Ex. state(icl>situation), state(icl>govt.).

Extra UW's	Denotes concepts not found in English. Ex. Soufflé(icl>food, pof>egg).
Restricted UW's	Set of Binary Relations grouped together to express a concept.

3.1.1. Example of a UNL Expression:-

Only a few farmers could use information technology in the early 1990's.

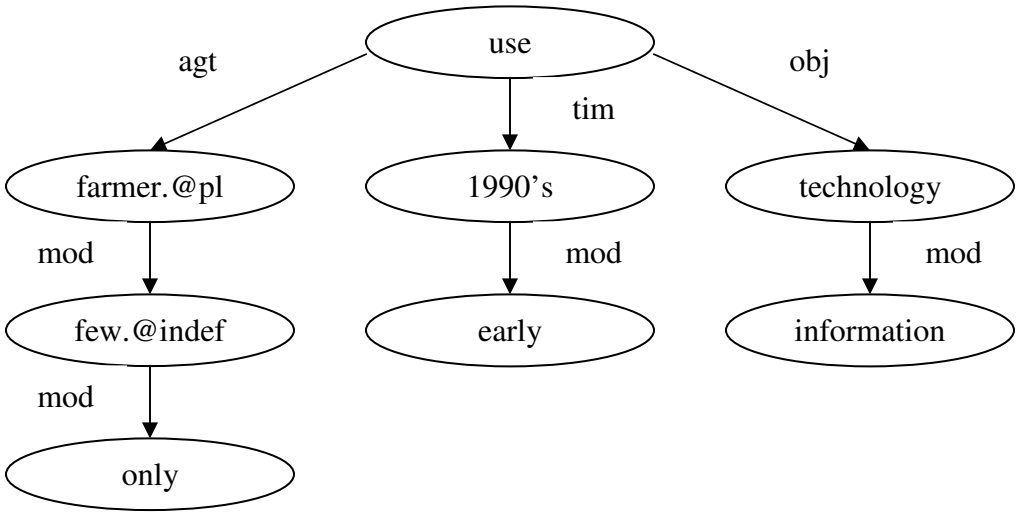
Core Sentence: Farmers use technology.

Specific modifiers are used to enhance this sentence so that it assumes its given form.

Equivalent UNL Expression:

agt(use(icl>do).@ability-past, farmer(icl>person).@pl)	…farmers use
obj(use(icl>do), technology(icl>thing))	…use technology
mod:01(farmer(icl>person), few(icl>number).@indef)	…a few farmers
mod(:01, only)	…only a few farmers
mod(technology(icl>thing), information)	…information technology
mod:02(1990's(icl>time), early)	…early 1990's
tim(use(icl>do), :02)	…use in early 1990's

3.1.2 UNL graph for the example:



Concepts are nodes and Relations are arcs. The Root of the graph is the entry node

Fig. 1: UNL Graph

4. Our work on the Pre-processing and Searching Modules:-

We have developed a meaning based search engine specific to the agricultural domain called “Agro-Explorer”. It uses UNL for intermediate representation. This not only makes it language independent but also lets us perform meaning based searches.

We have used C++ programming language, Shell scripts for the backend, HTML and Java Script for the front end and PHP for server side scripting in implementing the project. We also use, ENCO, a software that converts the English language query into its equivalent UNL representation.

The project was implemented on Linux. ENCO runs on the Windows platform. To overcome this difficulty we made an Enconverter server on Windows machine. This server had ENCO as its backend and when sent a query in English language it replies back with the UNL equivalent of the query. Whenever the search engine (running on Linux) requires an en-conversion, it sends requests to the Enconverter server and gets back the UNL version.

The Crawler module is used for crawling the web for all agricultural specific documents and making a corpus. This module is still to be implemented. We presently use a manually collected corpus for our searches.

In the Enconverter module, the natural language documents are converted into their equivalent UNL representation. This can be accomplished by using the Enconversion software called “Enco”. At the moment, we already have the corresponding UNL representation for our corpus. This module has been implemented and is used for converting natural language queries. This module runs on a Windows platform.

In the Pre-processing module, the corpus is cleaned and all the compound Universal Words (UW’s) are converted into an intermediate representation, which the search engine uses.

We have designed two user interfaces for the input module. The first interface accepts the query in its UNL representation itself. This UNL query is sent for Pre-processing. The processed query is then sent to the Search module where it is searched in the UNL corpus of all documents.

When a match is found, it writes the name, the original language, the relative links to the original document, the English translation and the UNL representation of the same, into a file. It also calculates a Relevance Score for each matched document. This Relevance Score is calculated by normalizing, the number of times the query occurs in the document, by the number of sentences in the document and multiplying the result by 100. This is also written in the file. Finally, this file is sorted according to each document’s Relevance Score. On the basis of this sorting a sequential list of

the results is formed. The file is then read by the PHP script and displayed in the sorted order. The first line gives the name and the link to the original language document. The next line gives the original language. We then display the links to the English language and UNL conversions of the document. If the original language of the document is English, the English language link is not displayed. The Relevance Score of the document is also displayed [2], [3].

The second interface accepts the query in English. This English query is sent to the Enconversion module. The resulting UNL query is tackled as described above.

5. Strengths of Agro-Explorer:-

1. Agro-Explorer eliminates the language barrier. Language barrier is the biggest obstacle in taking knowledge to the masses. Generally all the information, especially on the web, is in English or other major world languages. The majority of the population of the world doesn't understand these languages. To make this information available to all, the information has to be made language independent. There are two methods of doing this

    - The first method is to convert every language into all other possible languages. But this is not feasible. If we consider that there are only 10 languages in the world, we will have ${}^{10}P_2 = 90$ translators.

    - The second method is to consider an intermediate language. Universal Networking Language (UNL) is one such language. All the documents are converted into this intermediate language and a document can be converted back into any other language, from this intermediate representation. This makes the method extremely compact. For 10 languages, we need only $10 \times 2 = 20$ translators.

    We use the second method in our project. We are originally considering only four languages namely English, Hindi, Marathi and Konkani. When a user submits a query, he has to choose the language of his choice. He gets the results displayed in the language he chooses.

2. Agro-Explorer is a Meaning Based Search Engine. UNL, the intermediate language, uses Meaning Representation, i.e. it not only stores a word but also its meaning and attributes. For example, a word like "drink" will have different meanings in different sentences. It might mean "putting liquids in the mouth", or "liquids that are put in the mouth", or "liquids with alcohol", or "absorb" etc. But a UNL representation "drink(icl>do,obj>liquid)" denotes the subset of these concepts, "putting liquids into the mouth" which in turn corresponds to "drink", "gulp", "chug" and "slurp" in English. Attributes of a word provide information about how these concepts are being used in a particular sentence.

   First, all the documents are converted into their UNL representation. When the user enters his query, it is also converted into its equivalent UNL expression. Then, this query is searched in the corpus. As both the documents and queries are in their UNL representations, they are unambiguous and only documents that exactly match the query are retrieved.

   The results are much more accurate than any other techniques currently used. There are a few Meaning Based search engines on the web like www.exite.com [5], www.oingo.com [6] and www.simpli.com [7] but their results are vague. Universally recognized as the best search engine, www.google.com [2], is not a meaning based search engine. It uses text matching techniques. Its results are highly relevant but it returns a lot of extraneous pages that are not related to the searched query. For example, for a query "Prime Minister of India", its initial results are accurate but the documents at the end have only partial matches, i.e. they might only have "Prime" or "Minister" or "India" or some combination of these words. This not only takes much more time but also wastes storage space and hogs the bandwidth.

   In Google, we can specify that only exact matches be retrieved. Hence, only pages that have "Prime Minister of India" as a phrase will be retrieved. But pages that have "Indian Prime Minister" will not be retrieved. In our case both these cases will be matched.

6. Suggested improvements in Agro-Explorer:-

1. The use of Agro-Explorer is limited by the level of literacy in rural India. It can only be used by individuals who have basic reading and writing skills, not necessarily in English but in any of the supported languages. To make it even more accessible, an icon/picture based server will be very helpful.

2. A question and answer facility will greatly enhance its acceptance for the rural masses. A semi-literate person may find it difficult to extract the information he needs, even from highly accurate results provided by the search engine. To overcome this problem, it can be interfaced with the Q & A site project under Dr. K. Ramamritham.

3. ENCO, the software used to convert language documents into UNL, is not fully developed; a few difficulties might be encountered while trying to convert some sentences. This problem will be solved as more research is done on its lexicons and the rule base.

4. A crawler has to be designed that will crawl the web and retrieve all documents pertaining to the Agricultural domain, irrespective of their language and convert all the documents into their corresponding UNL representation.

7. References:-

1. www.unl.ias.unu.edu
2. www.google.com
3. www.markhorrell.com/seo/pagerank.html, Google's ranking algorithm.
4. www.google.com/press/guide/index.html, Google's ranking algorithm.
5. www.exite.com
6. www.oingo.com
7. www.simpli.com (not working presently)

8. Acknowledgements:-

We would like to extend our gratitude towards Prof. P. Bhattacharya, for his able guidance throughout the course of this project. The successful completion of the project would have been an uphill task without his attention, motivation and help.

We also thank the staff of the TCS Lab, Deptt of Computer Science and Engineering, I.I.T Bombay, for their continuous cooperation. Working with Anand Shinde (System Administrator), Jagadish Khot and Mrugank S. Surve made working in TCS Lab a pleasant experience.

The support extended by the members of Industrial Design Center (IDC), I.I.T Bombay, imbued in us the urge to improve our search engine and make it more user-friendly.

Equally praiseworthy was the work of the staff of Media Lab Asia, Mumbai hub, in maintaining good co-ordination between the different people associated with this project.

Sarvjeet Singh
Tushar Chandra
Upmanyu Misra
Ushhan D. Gundevia

Date: 15th July, 2002.