[ICUKL November 2002, Goa, India]

A Multi-Lingual, Meaning Based Search Engine

Dr. Pushpak Bhattacharya[a], Sarvjeet Singh[b], Tushar Chandra[c] , Upmanyu Misra[c], Ushhan D. Gundevia[c]
[a] Prof., Computer Science and Engg. Dept., I.I.T. Bombay. pb@cse.iitb.ac.in
[b] Final Year B. Tech. student, Computer Science and Engg. Dept., I.I.T. Bombay. sarvjeet@cse.iitb.ac.in
[c] Final Year B. Tech. student, Computer Science and Engg. Dept., I.E.T. Kanpur.
{ tushar_chandra, upmanyu_m, ushhan }@rediffmail.com

Abstract:

In this paper, we present a meaning based search engine that can be used as a multi-lingual platform for all sorts of search queries. We have used the Universal Networking Language (UNL) as the underlying communicator. We try to surpass the language barrier at the World Wide Web (WWW) level. WWW is the largest repository of knowledge known and a language gap here is obviously a big drawback. Although, we strongly believe that this hiatus is surmountable and the search engine is an early effort in this direction.

We discuss the need to develop a meaning based, inter-lingual search engine and its underlying principles. A brief discussion of UNL, the underlying intermediate language, is also given. We present the model, explaining its various modules and their implementations. Furthermore, the strengths of our search methodology, with respect to other techniques, are highlighted.

Keywords:

Multi-linguality, Sense Disambiguation, Linguistic Issues, UNL, WWW, Encoder, Decoder.

1. Introduction:

Internet plays an important, and at times vital role, in the day-to-day functioning of our life. The vastness of knowledge available on the WWW is the cause of our ever-increasing vulnerability to “not the best” knowledge available. As is known, search engines are the largest contributors towards the mining of knowledge from the Internet. With the steadily growing power and reliability of Natural Language Processing, the UNL [1] can be a generous contributor in the realization of highly dependable search engines.

In this paper, we present a search engine that attempts to answer the question: “What and where is the best knowledge that a user seeks through his queries?” First, the reach of a search engine is expanded by making it multi-lingual. Thereafter, the “prospective” results are searched in the UNL database that has been accumulated using crawlers and a natural language to UNL encoder software. Finally, the “best” results from the “prospective” ones are sorted by the meaning based behavior of the search engine. The model we present, thus, is ex-ante, “prescriptive”, and “preventive” in nature. To provide a systematic and comprehensive formulation of user requirements and preferences a ranking algorithm has been employed. The complete functioning of the search engine can be schematically represented as in Fig. (3).

This model retrieves all and only the relevant knowledge ranked according to its utility. We have tried to bridge the language gap by using an underlying, structured language as a backhand translator. As far as we know, we are the first to employ this technique.

2. Review of Literature:

In this section, we will briefly review the literature related to our software. This allows us to put our model in the perspective of the field.

a) Existing Search Engines:

The most common search engine on the web, Google, is widely believed to be the best example of a traditional search engine with is restricted to a single language (English in this case). Also, it returns a lot of useless and sometimes garbage information, which not only take up a lot of computing and bandwidth resources, but also are very cumbersome. It uses the famous random walk algorithm, which ranks the documents according to the link structure, coupled with the local query specific score to give the final rank to a page. [2]

b) Existing Meaning-Based Search Engines:

Only a few meaning based search engines have been developed so far. Though the attempts made are highly laudable but their results are nowhere close to the mark and therefore they failed to be commercial successes. A few of such earlier attempts are mentioned below.

Search engines like oingo.com, excite.com and simpli.com also provide meaning based searching. Launched in October 1999, Oingo has already introduced three fully functional products: DirectSearch, DomainSense and AdSense. DirectSearch, a meaning-based search technology, uses the company's ontology to provide more precise and effective search results. DomainSense, Oingo's meaning-based domain name suggestion technology, currently increases domain name sales for leading registrars around the world. AdSense serves the most highly targeted advertisements on the Internet; effectively targeting advertisements based on search meanings rather than keywords.

Dmitry Sergeevich Ermolaev, a Russian programmer, invented an "Intelligent Semantic, Clever, meaning-based (search engine) Searching System for Text Information"[3] which took into consideration the meaning of specific words or queries being searched. Unfortunately, this could not gain much industrial popularity.

Another step in the direction of meaning based searching was taken by a project of Chinese Academy of Sciences. This project [4], accomplished in 1998, worked on simplified Chinese and English. This project provided the flexibility of adding Traditional Chinese (Big5) and Traditional Chinese (EUC) in the future. Its established system consisted of two subsystems: Organization based subsystem and Web page based subsystem. Organization based subsystem was developed specially for users to find whether a certain organization in China had its own website and some detailed information about that organization. The web page based subsystem was developed for users for searching information in the web documents.

c) Existing Multilingual Search and Information Retrieval resources:

A Multilingual Information Retrieval Tool Hierarchy (MIRTH) [5] for the World Wide Web gave a general model of multilingual information retrieval for Web searching. It coped with both English and Chinese information retrieval. MIRTH first created an index file that contained key information about different Web pages. MIRTH indexed both document titles and document contents. Users could key in queries of search terms directly via a Web browser. Then, MIRTH started a search program that explored the pre-computed index files in real time and yielded search results accordingly.

“MULINEX: Multilingual Web Search and Navigation Tool” [6], developed by a consortium consisting of five European companies, supports English, French and German. It supports selective information access, navigation and browsing in a multilingual environment. The project emphasizes a user-friendly interface, which supports the user by presenting search results along with information about language, thematic category, automatically generated summaries, and allows the user to sort results by multiple criteria.

A comprehensive review of the work done in the field of multilingual information management can be found in a report titled “Multilingual Information Management: Current Levels and Future Abilities” [7]. A site containing links to 2 language multilingual search engines can be found at [8]. The European Commission has been working very hard to promote multilinguality, especially within EU in its V EU Framework Programme.

d) Universal Networking Language:

UNL [1] is 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. In this subsection, we are providing a brief description of UNL semantics and the way it works.

Binary relations are the building blocks for UNL documents. They are made up of two Universal Words (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.

UWs 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 UWs

Basic UW Character strings corresponding to English words.
Ex. Liquid, state.

Restricted UW Denotes a specific concept.
Ex. state(icl>situation), state(icl>govt.).

Extra UW Denotes concepts not found in English.
Ex. Soufflé(icl>food, pof>egg).

Compound UW Set of Binary Relations grouped together to express a concept.

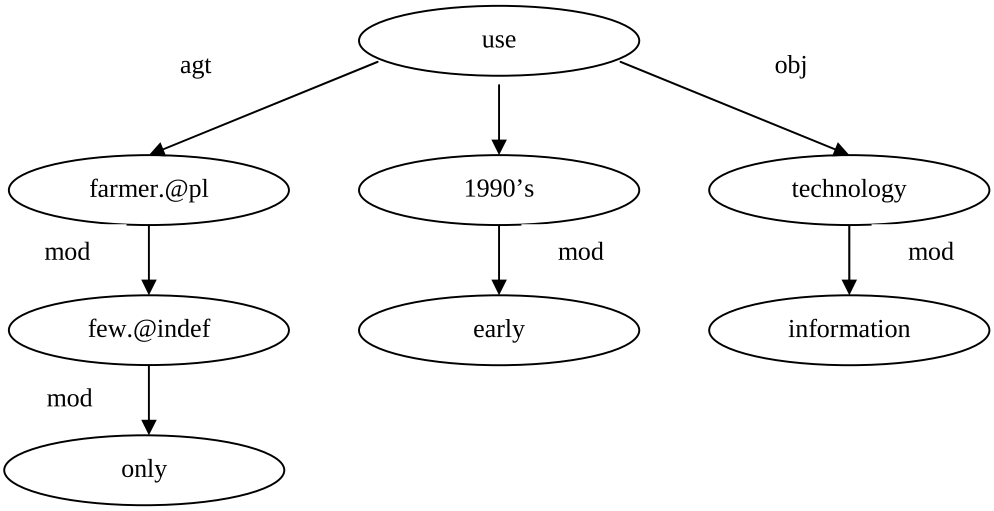
We give here an example of a UNL expression to provide an insight of the formulation.

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



Concepts are nodes and Relations are the arcs. The Root of the Graph is the Entry Node.
Fig. 1: UNL Graph

e) Encoder and Decoder:

To use UNL, we use software that en-converts the natural language into UNL and software that de-converts UNL into a natural language. Presently Encoders and decoder for English, Hindi, Spanish, Russian and Italian are under various stages of development.. The UNL system is shown in Fig. 2.

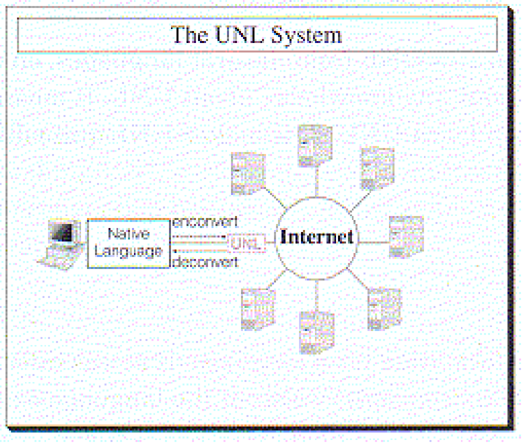


Fig. 2

3. Our Search Engine Model:

We have developed a meaning based search. 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, software that converts the English language query into its equivalent UNL representation.

The search engine is divided into different modules. It is explained pictographically in Fig. 3. They are as follows

a) Crawler Module:

The Crawler module is used for crawling the web for all documents and making a corpus. For crawling we plan to use the technique designed by Dr. Soumen Chakrabarti in his paper “Focused crawling: a new approach to topic-specific Web resource discovery” [9].

b) Enconverter Module:

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”. 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.

c) Preprocessor Module:

This module consists of a program and many helper/formatting shell scripts. The program takes as input, a UNL sentence, and outputs an intermediate representation, which is used by the actual search engine. The search has to be performed in a UNL database where documents are stored in a specific structure by the Encoder. This requires the query to be stored in a similar structure to that of the UNL database. This, again, in turn calls for an indexing for the documents in the database.

d) Search Module:

This is the nucleus of the engine. It takes in the intermediate data set generated by the preprocessing module and a query and performs search of the data set. It outputs the documents matched in decreasing order of their relevance along with lines matched and their relevance ratio.

e) Linker Module:

This module links up the search module to the interface. It consists of cgi-scripts, which are called by the interface module to get the search results. The interface module never interacts with the search module directly but only though these scripts.

f) Interfaces:

This module is responsible for interaction with the user and giving him back the results in a user-friendly way. This module is done entirely in HTML and PHP. We have designed two user interfaces for the input module. The first interface accepts the query in its UNL representation itself while the other take in the query in English language, which is then converted into UNL by the Enconverter Module.

When a user enters his query (either directly in UNL or English language, which is then converted into UNL), it is sent for preprocessing. 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, and 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. The Relevance Score of the document is also displayed.

4. Strengths of our Search Engine:

1. Our search engine 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 $10\times2 = 20$ translators.

We use the second method in our project. We are originally considering only English. If we use more languages (we could use any language for which Enconverters and Deconverters are available), he could to choose the language of his choice. He would get the results displayed in the language he chooses.

2. Our search engine 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, www.oingo.com and www.simpli.com but their results are vague. Universally recognized as the best search engine, www.google.com, 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.

5. Conclusion:

In this paper, we present a different approach to the problem of meaning based search engine and multi-linguality. We believe that making searches on the web meaning based is becoming the need of the hour. Also, the importance of multi-linguality should at least be at par with the other aspects of a search engine. For the search results to be realistic and meaningful, they must encompass the typical user's requirements and specifications.

The model in this paper is an amalgamation of two independent features. We integrated the user's language requirement with the relative importance of knowledge the user seeks. This has been possible by using the UNL as an intermediary language. UNL representation is language independent and captures the relationship between the words and their attributes. Hence multi-lingual and meaning based properties can be incorporated together rather than using separate language translators in search engines.

The bottleneck created by the exponentially expanding content of the web, more so in dominant languages, has been drawing attention to the area of sense disambiguation and linguistic issues, due to which the demand of producing more rigorous solutions, and hence more complex ones, has been constantly rising. Our model is a unique and early attempt to address this issue.

References

[1] The Universal Networking Language Specifications, Version 3, Edition 1, UNL Center, UNDL Foundation, www.unl.ias.unu.edu/unlsys/unl/UNL%20Specifications.htm

[2] “The anatomy of a large-scale hypertextual Web search engine”, Sergey Brin and Lawrence Page, Computer Networks and ISDN Systems 30 (1 –7), 1998.

[3] “My Invention—Intelligent Semantic, Clever, meaning based Searching System”—Dmitry S. Ermolaev.

[4] “A Multilingual (Chinese, English) Indexing, Retrieval, Searching Search Engine”, Wen-hui Zang, Wei Mao, Hua-lin Qian.

[5] MIRTH --Chinese/English Search Engine: A Multilingual Information Retrieval Tool Hierarchy For World Wide Web ‘Virtual Corpus’ and Training Resource in Computing and Linguistics & Literature, Thesis by Xioda Zhang, Leeds, UK, 1996.

[6] “MULINEX: Multilingual Indexing, Navigation and Editing Extensions for the World Wide Web”, Erbach, G., Neumann, G., and Uszkoreit, H. 1997, published in Hull, D. and Oard, D. eds. *Cross-Language Text and Speech Retrieval - Papers from the 1997 AAAI Spring Symposium.* AAAI Press, Stanford.

[7] *Linguistica Computazionale*, Volume XIV-XV, “Multilingual Information Management: Current Levels and Future Abilities”, Publisher: Insituti Editoriali e Poligrafici Internazionali, Pisa, Italy, 2001.

[8] Link to 2 language multilingual search engines, www.themillenniumsearch.com/

[9] “Focused crawling: a new approach to topic-specific Web resource discovery”, Soumen Chakrabarti, Martin van den Berg and Byron Dom, published in *Elsevier Science B.V.*, 1999.

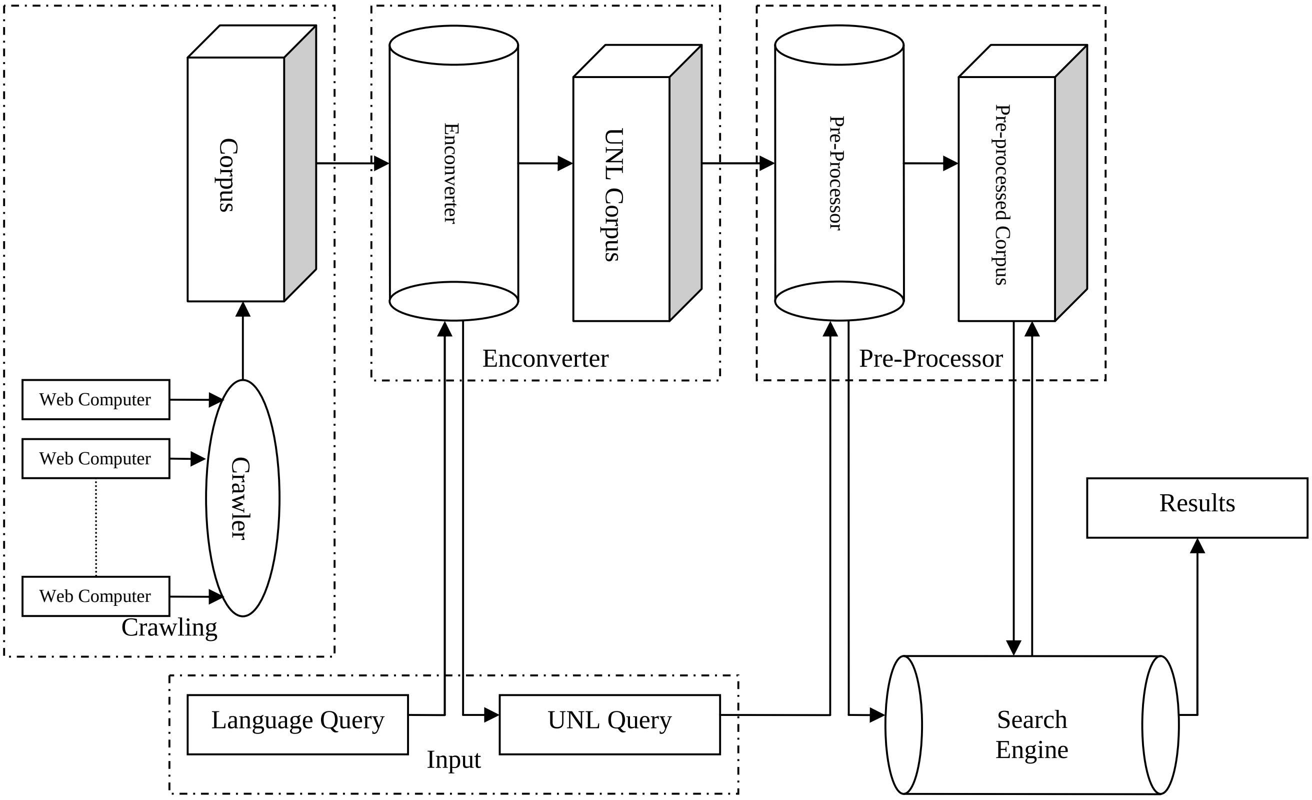


Fig. 3: Block Diagram of Agro-