CoRee – The UNL based Semantic Search

Balaji J, Umamaheswari E, Subalalitha C N, Elanchezhiyan K, Madhan Karky V,
Ranjani Parthasarathi, Geetha T V
Tamil Computing Lab,
Department of Computer Science and Engineering &
Department of Information Science and Technology,
College of Engineering, Guindy,
Anna University, Chennai -600025

Abstract:- ***Semantic search has become an active area of research in recent years. This has emerged from the necessity of providing semantically relevant information to the user, from the deluge of documents available on the web. A number of techniques have been proposed to extract the semantic content of documents which include use of wordNet, ontologies etc to identify related concepts. In this paper, a new framework for a semantic search engine, named CoRee, based on Universal Networking Language (UNL) is presented. UNL is a language-independent, concept based representation that captures the semantic structure of the documents. The significance of this work is that it incorporates the UNL based semantics in every component of a search engine namely, document representation, index, query handling, searching, ranking and presentation of the results.***

Although, the design of the concept based framework, CoRee, is generic and language independent, it has a language dependent component to convert documents to the UNL representation. Hence, this paper presents the design and evaluation of CoRee as a concept based search for Tamil Language. An evaluation of the CoRee framework for Tamil, for a corpus of 33000 documents of the tourism domain shows an improvement of 50% in the precision of the retrieved documents when compared with keyword based search.

1. Introduction

Today, there is a tremendous increase in the number of Web documents, not only in English but also in other languages such as Chinese, Arabic and Indian Languages. This information overload has resulted in serious research in the area of Web based search. Focus of the research is in tailoring both the offline and online components of the search engine to obtain relevant results. Most heavily used search engines such as Google, Yahoo and MSN primarily use keyword based search strategies [1]. The documents obtained from the search process are ranked by the query independent PageRank algorithm [2] or the query dependent HITS algorithm [3] which score documents based on incoming and outgoing hyperlinks associated with the web pages. However, most of these keyword based search engines produce a huge number of documents. The ultimate challenge of search engines is to produce relevant ranked information that exactly satisfies the users' information need. This challenge is further compounded by the vocabulary problem

arising from polysemy and synonymy. In this context, concept based search attempts to improve search effectiveness by incorporating semantic information rather than using the presence or absence of keywords as the basis for the retrieval process.

Concept based search mechanisms can be classified as those that use background knowledge source to provide conceptual information and those that use semantically analyzed components of the document or a combination of both. Concept based search can also be classified based on how the semantics is used to represent the documents. Documents can be represented by considering concepts associated with the frequently occurring keywords or by converting important components of the document into a semantic structure [4]. In addition, concept based search can also be classified based on where the semantics is introduced in the components of the search engine [4]. Semantics can be introduced in building the index, query expansion, searching and also in ranking the search results.

This paper describes a concept based search engine, CoRee, which uses semantic representation of the sentences of the documents, and incorporates semantics in all the components of the search engine. Universal Networking Language (UNL) [5], an inter-lingual, language independent, semantic representation is used for document representation. Universal Networking Language (UNL) was originally designed to aid machine translation [6]. UNL deals with concepts, that are represented as Universal words (UWs) and defines a set of relations that can exist between them. A sentence in natural language is represented by UNL as a directed graph in which nodes represent concepts and links represent relations [6]. In addition attributes are also associated with both concepts and relations of the UNL graph. Each UNL concept of the graph is obtained by referring to a UNL knowledge base (KB) which provides concepts for terms/words of the natural language. Each concept is also placed in a UNL specified concept hierarchy represented as constraints in the UNL KB. The UNL consortium has at present defined a standard set of 46 UNL relations which essentially represent the semantic relations between concepts of the sentence.

The UNL based semantic representation of the sentence has been used for a number of language technology applications. Although originally designed for machine translation, UNL representation has also been used for semantic based text summarization [7] and to build concept based indexes for multilingual search engines [8].

This paper outlines the architecture of CoRee, a multilingual conceptual search engine powered by UNL. It explores the use of UNL for building a richer conceptual index which facilitates multilevel searching and ranking. In addition, it analyses the UNL based conceptual index and uses it for effective context based query expansion. Although the architecture of CoRee is generic, this paper describes the UNL concept based search of Tamil documents from tourism domain and is part of a project on Cross Lingual Information Access funded by the Ministry of Information and Communication, Government of India.

The paper is organized as follows: Section 2 discusses related work in the area of conceptual

and semantic search. The architecture of CoRee is described in Section 3. Sections 4, 5, 6, 7 and 8 respectively outline UNL Enconversion, anaphora resolution, conceptual indexing, and context based query expansion and the conceptual searching and ranking components of CoRee. Section 9 presents conclusion and directions for further research.

2. Related work

A brief overview of existing approaches to semantic search is presented in this section.

2.1 Semantic Search Engines

In general, web based search uses a bag of words approach to represent documents [9]. On the other hand, semantic search engines associate concepts or semantic structure with words of the document resulting in a bag of concepts approach [10].

While some meaning based search engines use sentence level semantics, others use ontology as the background knowledge source for providing semantics. Hakia [11] is a semantic search engine that uses knowledge of Ontology and Fuzzy logic for semantic ranking. In order to retrieve conceptual results, it uses QDEX (Query Detection and Extraction) Indexing Architecture which enables semantic analysis of web pages and provides meaning based search results. On the other hand, SenseBot [12] is a semantic search engine that runs over search engines like Google and Yahoo to generate multi-document summary based on text mining and limited semantics.

Concept based search can also be based on the use of knowledge structures. One such search engine is Engineering/Environmental Knowledge Ontology-based Semantic Search EKOSS [13]. It uses a fully functional ontology for representing the knowledge base. It provides a collaborative knowledge sharing environment and helps knowledge experts to share their knowledge such as research papers, databases, computer simulated models and even curriculum vitae. The EKOSS system is used to construct computer-interpretable semantically rich statements of the knowledge resource. When a user request is posted, this system converts the user request into a computer readable knowledge description based on description logic and associated rules.

Ontology-based information retrieval [14] intended for e-Government has been developed for searching government legal documents. The disadvantage of using ontology based search engines is that they are susceptible to changes in the information resources. More over, the effort required to build ontology is huge. This task is domain dependent and the use of common vocabulary ontology for different domains remains a challenging task.

2.2 UNL based Search Engines

A meaning based multilingual search engine that uses UNL (Universal Networking Language) is the AgroExplorer [8]. This search engine is similar to the CoRee search engine described in this paper, since AgroExplorer also uses Universal Networking Language (UNL) expressions for representing sentences as graphs that capture the meanings of the sentences. The system has been developed for agriculture domain and also provides multilingual feature. It uses a simple search and rank process based on the degree of match of the query UNL and the frequency of occurrence of the concepts with other concepts in the UNL expression. CoRee, the concept based search engine discussed in this paper, incorporates semantics in every

component of the search engine. A richer UNL based conceptual index has been used, and in addition the conceptual index has been used for context-based query expansion. The searching and ranking used in CoRee differs from that used in AgroExplorer in that CoRee uses a sophisticated three-level search and rank process based on degree of match, the relations between concepts and the conceptual index based query expanded concepts. The detailed description of the basic components of CoRee is given below.

3. The Architecture of CoRee

The overall architecture of CoRee, the concept based search engine is given in Figure 1.

CoRee consists of offline and online components that cater to UNL based conceptual search. The offline process begins with UNL enconversion of the documents obtained from a focussed crawler. To facilitate the enconversion, complex sentences are split into sentence constituents and enconverted. The purpose of UNL enconversion is to build a conceptual index. The enconversion process is enabled by the rich morphology of Tamil to be based on features of the words and their preceding and succeeding context only. In Tamil, the use of case relation indicated by morphological suffixes, POS tag and word level semantics allow the rule based enconversion to be independent of the syntactic structure of the sentence. The enconverter incorporates modules for anaphora resolution, word sense disambiguation and for handling multi-word expressions. The UNL graphs represented as multi-list structures are used to build the conceptual based index and also serve as the basis for the Template based Summarizer. The indexer identifies isomorphic sub-graphs to build three distinct indices – Concept (C), Concept-Relation (CR) and Concept-Relation-Concept (CRC). The UNL index besides being used for concept based searching and ranking, is also used to extract a snippet from each document.

During the online process, a user given query is first translated to a UNL graph using a light weight enconverter. Concept associations obtained from the UNL index are used for context based query expansion. The resultant set of UNL query graphs are given to the search and rank module which matches query graphs with the three conceptual indices based on degree of match, the relations between concepts and the conceptual index based expanded concepts. The ranked documents along with the associated snippet and summary are presented to the user.

The following sections describe the existing approaches in the design of the major components of CoRee and discuss the challenges handled, methodologies adopted and the evaluations carried out.

4. UNL Enconversion

4.1 Existing Approaches to UNL Enconversion

The UNL Enconversion is basically a semantic analysis problem where various linguistic features are used based on the characteristics of the input natural language. The Enconversion process can be classified based on the type and degree of usage of various linguistic features. The use of linguistic features such as morphology, syntax and semantics for the Enconversion depends on how the various UNL relations are characterized by the specific natural language.

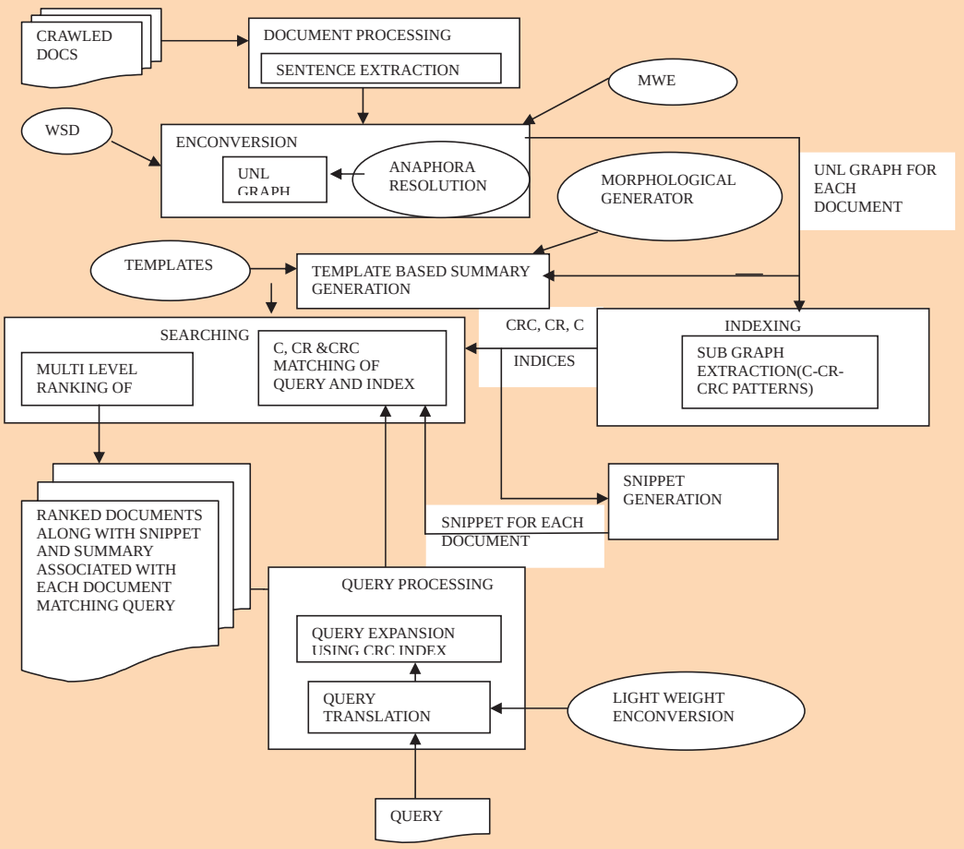


Fig. 1 Architecture of CoRee

An English UNL Enconversion described in [15] is centred on the use of a standard syntactic parser with a limited amount of morphology and semantics. A French UNL Enconverter [16] generates UNL expressions using an incremental parser which converts the expressions into a semantic graph using a rule based approach. Enconversion using a multi functional linguistic processor ETAP-3 [17] converts Natural Language into ETAP-3's internal representation that is essentially a normalized syntactic structure (Norm SS) which is then converted to the required UNL representation. Again, this approach is centred on the use of syntactic structures to aid the Enconversion process. Enconversion process for Arabic, a highly inflected language of relatively free word nature uses a strong interaction between morphological and syntactic processing modules. Therefore, a rule based approach for Arabic Enconversion cannot use the normal pipeline model where morphological analysis is followed by syntactic analysis for resolving ambiguity [18].

Another approach for English to UNL Enconversion [19] uses a two stage process where the conceptual arguments are first identified in the form of semantically relatable sequences (SRS) which are potential candidates for being linked by semantic relations. These are then mapped to form a parsed output and then UNL expressions [19]. Most of the Enconverters discussed above use a rule based approach. A statistical approach using a parser with associated morphological and syntactic linguistic features has also been attempted [20]. Most of the approaches discussed above depend on the syntactic structure of the sentence of the language, where the language in most cases has fixed word order characteristic.

Many Indian languages do not have a rigid fixed word order structure and the amount of freedom allowed by a language differs based on the richness of the morphology. In languages like Hindi and Bangla for which enconversion has been done, isolated case markers along with local word grouping, generally specify the case relations between the different components of the sentence. Since most UNL relations are based on case roles, the handling of this aspect differentiates the approach to Enconversion for Bangla [21] and Hindi [8]. An approach for Bangla words to UNL deals with the verb morphology and the use of syntax tree [22]. Another work on Bangla to UNL Enconversion [23] uses a set of morphological and semantic rules for enconversion and fits the rules to the enconverter framework obtained from UNDL foundation. All the UNL Enconverters discussed above use a structural syntactic parser for fixed word order languages and a dependency like parser for partial free word order languages.

The Enconversion component of CoRee for Tamil uses morphological features and word based semantic features with limited amount of context, without explicitly using syntactic structures.

4.2 Tamil to UNL Enconversion

As already mentioned, Tamil is a morphologically rich and partially free word order language. Normally, Tamil follows a subject-verb-object form. Here, case markers are conveyed by morphology of noun rather than separately as prepositions as in English or as karaks in Hindi. Therefore, morphological analysis of the word, in addition to providing part of speech also provides case relation information about the nouns. Verb morphology provides the tense, aspect, mood, person, number, and gender of the verb. Normally the person, number and gender of the verb are expected to match with the noun acting as subject/agent of the sentence. The richness of the noun morphology allows noun phrases indicating different cases to occur at any position in the sentence.

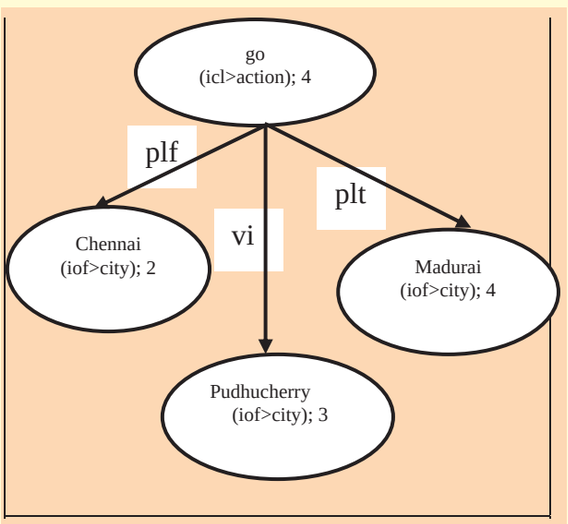
Information conveyed in other languages by position of sentence constituents and syntax, is conveyed by morphological features in Tamil. In CoRee, this characteristic of the Tamil language is exploited in the Tamil to UNL Enconversion process. As an example, a simple Tamil sentence is enconverted, and the resulting graph and associated UNL expressions are shown in ***Example 1***. In this example, morphological case suffixes such as ***ilirundhu*** signals multiple UNL relations such as "***frm", "plf",*** ***" tmf", "src",*** while the case suffix ***kku*** signals UNL relations such as" ***to", "plt", "tmt",*** and the connective vaziyaaka signals "*via*" relation. The relations are disambiguated using the semantic constraints of the words in which the case suffixes are attached. Thus in *Example 1*, the case suffix il irundhu associated with Chennai and "kku" associated with Madurai indicate the semantic constraints as iof>city which narrows down the relations of the verb with these two concepts to "*plf*" and "*plt*" respectively.

Example 1 :

Chennaiyilirundhu pudhuvai vaziyaaga

maduraikku sellalam

(சென்னையிலிருந்து புதுவை வழியாக மதுரைக்கு செல்லலாம்)




```
{unl}
{w}
Chennai (iof>city); 1
Pudhucherry (iof>city); 2
Madurai (iof>city); 3
sel go(icl>action); 4 ; @entry
{/w}
{r}
4       plf     1
4       via     2
4       plt     3
{/r}
{/unl}
```

In CoRee, complex Tamil sentences are broken down to simpler constituents for Enconversion. The Enconversion process takes place in two passes [24]. During the first pass, the possible UNL relations are identified using morphological or word features and stored in a relation vector. There are some UNL relations that can be unambiguously determined in this pass itself. However, as mentioned earlier, there are morphological suffixes which signal more than one UNL relation. In this context, the relation vector will have multiple values. During the second pass, the concepts that are linked by the UNL relation are determined. The second pass also helps in resolving ambiguity among the UNL relations obtained in pass 1. The disambiguation process uses the relation vector along with co-occurrence, parts of speech, connectives and semantics of the corresponding UNL concept and the context. The resultant graph obtained is stored in a multi-list structure. The details of these processes are given below.

4.2.1 UNL Relations Extracted in First Pass

In this pass, word based features required for UNL enconversion are obtained. The presence of a UNL relation is signalled by morphological

suffixes, adjectival suffixes and adverbial suffixes or by the presence of certain standalone connective words. This is explained below with suitable examples.

(a) **Case suffixes signalling one UNL relation**

When case suffixes are attached to the noun, certain UNL relations are signalled unambiguously. Most often, the noun to which the case suffix is attached is one of the words taking part in the UNL relation. The UNL relation obtained through the case suffixes of nouns are "***pos", "ben", "obj".*** This is shown in Example 2.

(b) Case suffixes signalling more than one UNL Relation

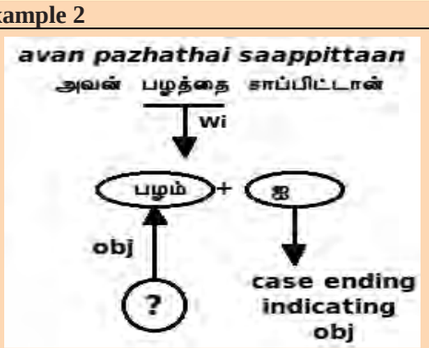
Certain case suffixes attached to nouns ambiguously signal more than one UNL relation. Suffixes of this category are given in Examples 3 and 4. The exact UNL relation will be determined during second pass using semantic constraints.

(c) **Adjectival suffixes and Adverbial suffixes**

The semantics "***mod"*** and "***man"*** can be determined unambiguously by the Adjectival suffix (***aana, iya***) and Adverbial suffix (***aaga***). Again the respective adjective or adverb forms one word connected by the UNL relations "***mod"*** or "***man"***. This is shown in Examples 5 and 6.

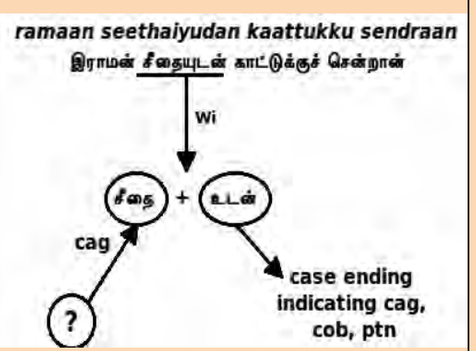
one word connected by the UNL relations "***mod"*** or "***man"***. This is shown in Examples 5 and 6.

Example 2



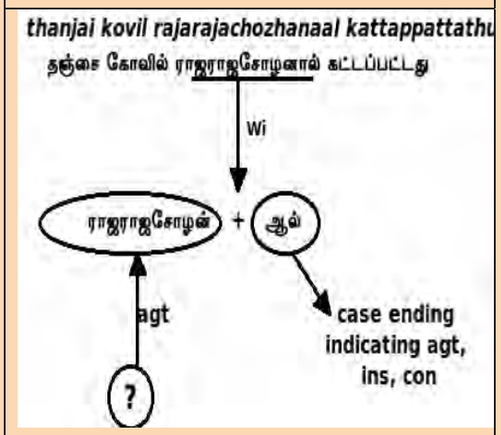
As shown in the associated figure**,** in the word pazhaththai (**Wi**) which has the case ending ***ai*** (Accusative case marker), the case ending indicates that **Wi** should be related to some other concept in the sentence through the UNL relation "***obj"***.
Similarly, the UNL relation, " ***ben"*** can be obtained by the case ending ***ukkaaga*** and "***pos"*** relation can be obtained by the case endings ***udaya, in*** and ***athu*** (Genetive case marker).

Example 3



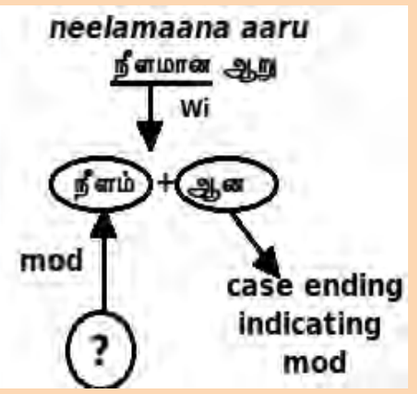
This ***example*** shows the case suffix ***udan*** or ***odu*** associated with word **Wi** indicating multiple UNL relations such as "***cag", "cob"*** and "***ptn".***

Example 4



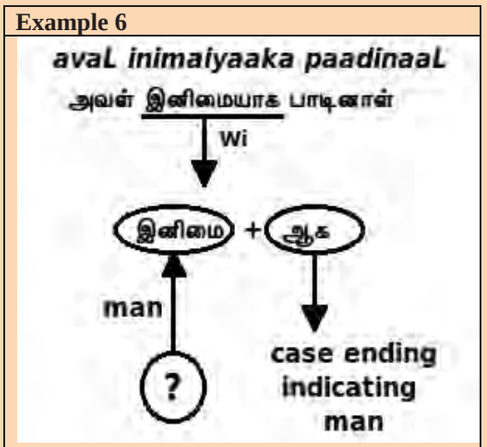
This example illustrates the signalling of the UNL relation "***agt", "ins"*** and "***con"*** for the case suffix ***aal*** associated with the word **Wi**.

Example 5



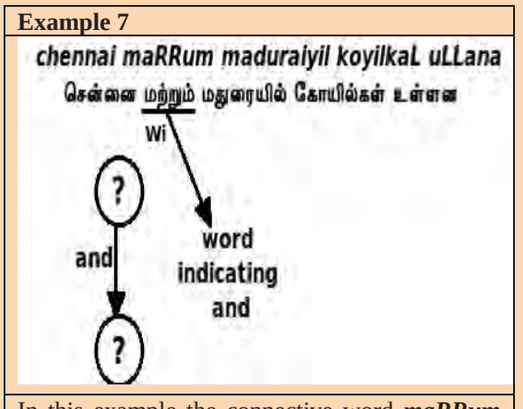
This example shows the signalling of "***mod"*** relation using adjectival suffixes.

Example 6



Signalling of the "***man"*** relation by the adverbial suffixes is shown here**.**

Example 7

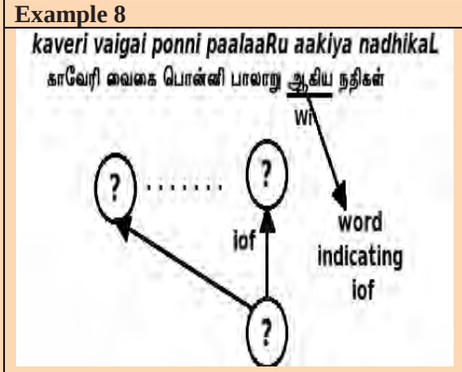


In this example the connective word ***maRRum*** indicates the UNL relation "***and***" between two different concepts.

(d) Presence of Connectives

Certain connectives signal certain semantic relations unambiguously. However, in these cases the connective word itself does not take part in the UNL relation. The connectives maRRum, mattumallaamal, aagiya, pondra, mudhaliya, vazhiyaaka, allathu, enpathu, and idaiye come under this category. Certain UNL relations such as "iof" and "nam" connect more than two concepts. This is shown in Examples 7 and 8.

Example 8



This example illustrates the use of the connective ***aagiya*** which connects more than two concepts.

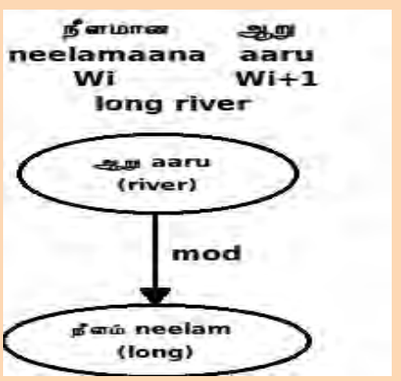
4.2.2 Second pass – UNL Graph construction

The second pass of the UNL enconverter performs two tasks, using additional information such as POS and the semantic category of the UNL relation signaling word $\mathbf{W}_i$, along with its context $\mathbf{W}_{i\pm k}$. In the context where UNL relation has been unambiguously determined by the first pass, there is a necessity to find the two concepts that are involved in the UNL relation in order to build the UNL graph. However, when there is an ambiguity regarding the UNL relation, this ambiguity has to be resolved before the UNL graph is constructed [24].

The UNL relation of nouns with case suffixes are usually with the corresponding main verb. The verb normally occurs at the end of the sentence. Therefore, if POS of $\mathbf{W}_{i+k}$ where (k>=1) word is a verb, then the $\mathbf{W}_{i+k}$ word is connected to the case suffixed noun $\mathbf{W}_i$ by the indicated UNL relation. In the case of adjectival and adverbial suffixes, the adjective or the adverb $\mathbf{W}_i$ is normally associated with the succeeding $\mathbf{W}_{i+k}$ whose POS is noun or verb respectively. This is illustrated in *Example 9.*

Linguistic features are used to resolve ambiguity of UNL relations. Possible UNL relations such as [***frm, plf, tmf, src***] are stored in a relation vector. The semantics of the word $\mathbf{W}_i$ (icl>time) is checked for obtaining the correct ***relation*** to the main verb of the sentence [24]. Example 10 shows disambiguation of the set of relations ***"cob***", "ptn", and ***"cag".***

Example 9

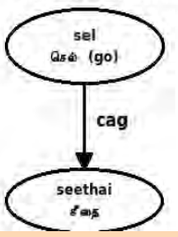


Example 9 illustrates the "***mod"*** relation obtained between the concepts. Here, the word $\mathbf{W}_i$ with the adjectival suffix ***aana*** is connected with the corresponding noun $\mathbf{W}_{i+1}$.

Example 10

இராமன்	சீதையுடன்	காட்டுக்குச்	சென்றான்
Raman	seethaiyudan	kaattukku	sendraan
Wi-1	Wi	Wi+1	Wi+2

raman went to forest with seethai



In this example UNL relation "***cag***" can be obtained using the semantics of the word $\mathbf{W_i}$ as (icl>person). Similarly, the UNL relations "***cob***", "***ptn***" can be obtained with the semantics of the word $\mathbf{W_i}$ such as icl>comrade, icl>place etc.

4.2.3 Evaluation of Enconversion

The enconversion process uses 53 rules encompassing the various categories of signalling relations. The UNL list and a multiword list have been used for extracting the concepts. The UNL List contains 29493 entries and multiword list contains 2249 concepts. The performance of the enconversion process has been evaluated using tourism domain corpus of 33000 documents.

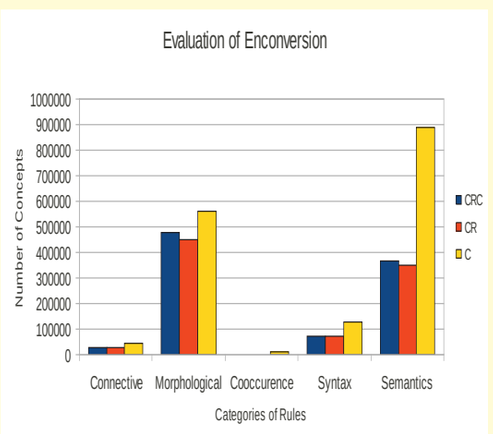


Fig 2 Evaluation of Enconversion

Fig. 2 shows the number of Concepts (C), Concept-Relation (CR) and Concept-Relation-Concept (CRC) obtained by the Enconversion process. It can be seen that maximum number of Cs and CRCs are obtained from morphological information and semantics. 90% of the enconversion is achieved using morphological rules and semantic rules. On performing human verification to determine the correctness of UNL, 80% of the enconversion was found to be satisfactory. On analyzing the 20% that was not correctly enconverted, it was found that the use of sentence constituents limited the accuracy of the enconversion. However, since this version of CoRee was designed to consider only simple UNL relation based indexing, the accuracy in sentence constituent enconversion was considered more important. Another aspect that was found to affect the indexing was the lack of connection between concepts of different sentences due to absence of anaphora resolution. Therefore anaphora resolution was added. The next section briefly explains the anaphora resolution module.

5. Anaphora Resolution in Enconversion

We first describe the existing approaches to anaphora resolution before describing the UNL based anaphora resolution.

5.1 Existing Approaches to Anaphora Resolution

The existing pronoun resolution approaches have mostly dealt with person pronouns. Commonly used approaches for anaphora resolution are based on the Centering theory proposed by Brennan et.al [25], and on the Hobbs algorithm [26]. Centering theory [27] [25] relates focus of attention, choice of referring expression, and perceived coherence of utterances within a discourse segment. In general, Hobbs algorithm [28][26] is a syntax-

based algorithm, which traverses the syntax-tree using breadth-first search, left-to-right and looks for an antecedent matching the pronoun in gender and number.

Anaphora resolution has been carried out for various languages such as Tamil [29][30], English [31], Hindi [32], Chinese [33], Spanish [34][35], Turkish [36] etc using the approaches mentioned above. They predominantly use syntactic information to resolve the person anaphors. Some work has also been done in English using semantic roles such as agent and patient to find the correct antecedent of a person pronoun [37][38]. However, assigning semantic roles using parser is a difficult task in the agglutinative languages like Tamil. In addition, most anaphora resolution systems have dealt with person pronouns and to a certain extent plural pronouns and used a syntactic based approach for event pronouns. However, in CoRee morphological suffixes, semantic constraints, and semantic structure of sentences are used for anaphora resolution.

5.2 UNL based Anaphora Resolution

UNL semantics has been utilized for anaphora resolution in three ways. Firstly, UNL specified semantics constraints have been used to classify pronouns as person, place and events. Secondly, semantic constraints associated with single words representing referring expressions have been used for anaphora resolution of person, place and simple cases of plural pronouns. In these two cases, no UNL enconversion is performed. The UNL has been used only to tag words with semantic attributes and semantic roles indicated by morphological suffixes, and with semantic constraints specified by the UNL KB. Thirdly, UNL enconverted sentences available as semantic graphs are used to tackle referring expressions consisting of multiple words such as event pronoun and certain types of plural pronouns. Thus, two approaches are used for anaphora resolution:

(i) Use of UNL based Semantics Integrated Centering Theory (SICT) for person and place pronouns.

(ii) Use of UNL graphs

A brief description of these two approaches is given below.

(a) Using SICT

Here we use the UNL based classification of pronouns to persons, places and events, to narrow down the type of anaphors and filter out the non-anaphoric expressions with respect to the type of anaphor identified. Only one class of anaphora is considered for further processing. Moreover, the non-referring entities are filtered out using not only the agreement features, but also using the semantic roles conveyed by morphological suffixes of the referring expressions. Finally, the correct antecedent of a pronoun is identified by applying the transition rules such as continue, retain, and smooth shift and rough shift as discussed in [25]. With this level of semantics included in centering theory, we find that we can effectively resolve person pronouns, place pronouns and to a certain extent plural pronouns.

(b) Use of UNL Graphs and SICT

The UNL graphs obtained for a sentence are used for anaphora resolution. Here, two techniques are used for resolving the anaphors. One is the use of UNL relations to extract concepts represented as antecedents, and the other is the use of UNL sub-graphs represented as antecedents for anaphora resolution.

- *Use of UNL relations to extract the antecedents of a pronoun*

  To find the antecedents using UNL relations, the relations are divided into two broad categories, namely *coordinating*

UNL relations and subordinating UNL relations.

- *Coordinating UNL Relations* are relations obtained for referring expressions that exactly match with the UNL relation obtained for anaphors. In this case, antecedents of a pronoun are identified using the UNL relations obtained for referring expressions in the preceding sentence and UNL relations obtained for pronoun in the succeeding sentence. The connected relations in both the preceding graph and the succeeding graphs are matched. If the concepts in the preceding and succeeding graphs are connected to transitive verbs, then the referring expression which is connected by the relation which matches with the relation connected to pronoun is selected as an antecedent of that anaphor.

- *Subordinating UNL Relations* are relations obtained for anaphors that depend on the UNL relations obtained for referring expressions. Certain UNL relations connected with the referring expressions in the previous utterance depend on the UNL relations connected to a pronoun. A concept connected by an "obj" relation in the previous utterance is the antecedent of a pronoun which is connected by an "agt" relation. The relations "ben-agt" and "plc-aoj" are handled in a similar manner.

- *Use of UNL sub-graphs to extract antecedents of a pronoun*

  Resolving plural pronouns representing multiple words as their antecedents and event pronouns representing verb phrases, clauses and segments of sentences is a difficult task. An approach proposed for resolving event anaphors [40] identifies inflection verbs as their antecedents by exploring various features such as positional, lexical and syntactical features. In our approach, we use UNL semantic constraints associated with the words, UNL attributes and UNL relations obtained between the concepts. The set of possible referring expressions are identified using the morphological suffixes and semantic constraints of verbs and its connected concepts.

5.3 Evaluation of Anaphora Resolution

The efficiency of the anaphora resolution system has been tested using 800 tourism domain specific documents. We have identified 1562 person pronouns (singular), 72 plural pronouns, 1256 place pronouns and 58 event pronouns. A comparison of these results with the original Centering Theory is shown in Fig. 3, using the F-measure as the metric. The parameters considered for F-measure calculation are, the total number of pronouns identified, number of pronouns resolved and number of pronouns correctly resolved.

$$\text{Precision} = \frac{\text{number of anaphora correctly resolved}}{\text{number of anaphora resolved}}$$

$$\text{Recall} = \frac{\text{number of anaphora correctly resolved}}{\text{number of anaphora present in the corpus}}$$

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The use of original centering theory gives a low F-measure. This is because the ordering of referring expressions by grammatical relation may fail to identify the correct antecedent in Tamil, a partial free word order language. It can be seen that the inclusion of word level semantics (i.e. Semantics Integrated Centering Theory - SICT) gives better F-score. In the UNL graph based approach, we can see that, the plural pronouns and event pronouns have been tackled using UNL graphs. The use of UNL

graphs significantly increases the F-measure for singular person pronouns to 0.77 and plural pronouns to 0.71. Here, we have considered only simple UNL graphs to resolve event pronouns and thus obtained the F-measure as 0.57. Better performance can be obtained for event pronouns by considering complex graphs with nested relations.

The inclusion of anaphora resolution in the document processing helps in better representation of the document and hence an improved index where the resolved pronouns contribute to increased frequency of occurrence of words.

6. Concept Based Semantic Indexing

6.1 Existing Approaches to Semantic Indexing

Semantic indexing techniques like latent semantic indexing [41] and concept based vector space models [42] try to find the underlying latent semantic structure in the document matching with the query using statistical techniques and mathematical calculations.

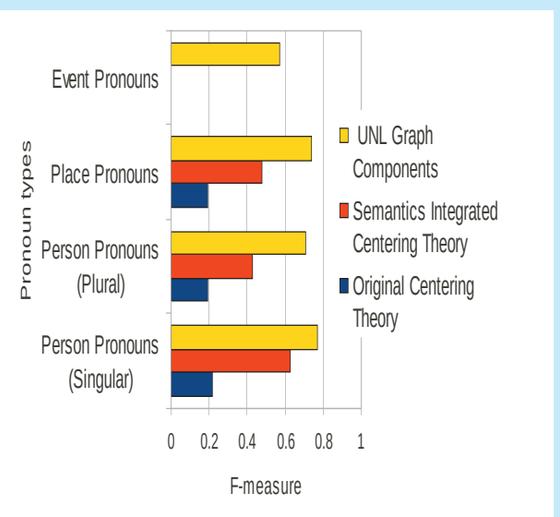


Fig 3: Evaluation of anaphora resolution

This may be time consuming and when done for large corpus may increase the offline processing time. A system that uses UNL for indexing is the Multilingual Meaning-based Search Engine – Agro Explorer. This system has done indexing at three levels, namely, UNL expressions (phrasal and sentential concepts), Universal Words (lexical concepts) and Keywords/Stem Words (Using Stemmers & Lucene). The indexer system is entrusted with the task of taking the UNL Corpus and generating an inverted index on it [43]. For CoRee, we have proposed an indexing mechanism that is designed to index concepts and their relations with other concepts when a UNL graph is given as input. The UNL indexer used by Coree not only focuses on indexing the terms/concepts [43], but also inherits all possible semantic features obtained from the UNL graph, resulting in effective search.

6.2 UNL Indexer

The UNL concept based indexing technique is designed such that the inherent semantics between the concepts in the document is retained by capturing the concepts along with their relations to other concepts. Since UNL documents are stored in the form of graphs, the basic properties of graphs are used to find the indices. Indexing large graphs proves to be a daunting task as the size of sub graphs that needs to be indexed is voluminous.

This section discusses the extraction of indices from UNL graphs represented as multi-list structures into appropriate conceptual indices by repeatedly extracting isomorphic sub graphs. The UNL indexer of CoRee [45] identifies three different sub graph patterns namely CRC, CR and C as indices. These indices inherit all the information maintained in the UNL graph such as Parts of Speech (POS) and weight factor of each index which yields its frequency. It

also maintains the document identifiers and the sentence identifiers of each index by using bit patterns. Sentence identifiers indicate the importance of the index in that particular document. This helps in ranking the document efficiently. The CRC pattern of UNL index is also used in query expansion which is explained in Section 7. Further the index information helps in building a snippet for each document, offline.

The CRC, CR and C type indices identified for the UNL graph shown in ***Example 1*** are given below.

C-R-C indices are:

Go (icl>do)-plf-Chennai(iof>city)

Go (icl>do)-plt-madurai(iof>city)

Go (icl>do)-via-pondichery(iof>city)

C-R indices are:

Go (icl>do)-plf

Go (icl>do)-plt

Go (icl>do)-via

C indices are:

Go (icl>do)

Chennai (iof>city)

Madurai (iof>city)

Pondicherry (iof>city)

These indices are stored in a Binary Search tree (BST). In order to avoid collisions and to add similar concepts of a term in the BST, two linked lists are connected to every node in the BST. The indices of a term are stored in one linked list represented as X_D1, X_D2 in Fig: 4 and similar concepts of a term are stored in another linker list, represented as X1, X2 ,....X_N in Fig 4. For instance, the nodes named X_D1, X_D2 hold the document identifiers of the UNL index inserted in the node X and the nodes X1, X2 ,....X_N will hold the concepts that denotes the other forms of X. For instance, if the concept chennai(iof>city) for the term, „சென்னை” is stored in node X then the chennai(iof>city) for the term, “மெட்ராஸ்” is stored in X1.

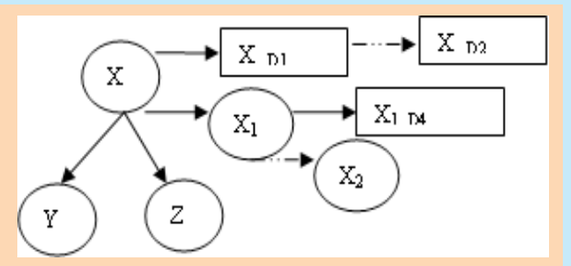


Figure 4: BST-Modified Node Structure

The data structure used for storing the index makes it efficient to be used in both online process like query expansion and offline process like snippet generation apart from search process. This makes the UNL indexer versatile catering to the various needs of the entire search system.

6.3 Evaluation of UNL Indexer

A performance analysis of the indexer has been carried out with 33000 Tamil tourism documents enconverted to UNL. Figure 5 shows the growth of size of indices of three types against the number of documents. It can be seen that, though there is a steady increase in the count of indices as the document count increases, the number of additional indices decreases. It can also be seen that the UNL indexer is scalable.

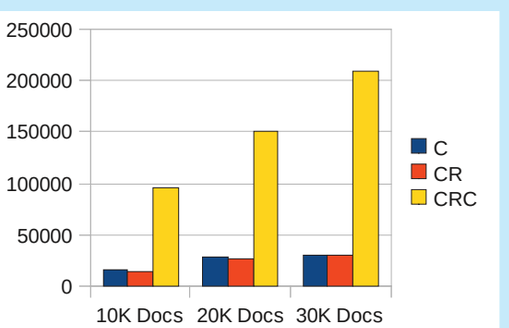


Figure 5: Size of indices versus Number of documents indexed

7. Context Based Query Expansion

7.1 Existing Approaches to Query Expansion for Semantic search

The semantic search engines discussed in section 2, use some form of semantic query expansion. In Hakia besides the keywords, phrases are used for meaning based searches [11]. The limitation of Hakia is that it accepts queries as questions in a specific format only. In addition, the QDEX algorithm that Hakia uses, extracts all possible queries that can be asked on the content of web pages of various lengths and forms. This is an offline process before any user query is entered. The major difficulty in QDEX system is the necessity to reduce the huge number of generated query sequences into a few dozens that make sense. Hakia allows only these predefined query sequences generated from the content to be used as queries.

Most search engines designed to provide meaning based results, require sophisticated query analysis techniques. Some search engines also use associated concepts from generic ontology for query expansion. The CRCs of the indexer is normally associated with a verb. However, search engine queries rarely use verbs. By associating query term concepts with CRCs from the indexer, we build a more effective query graph which includes verbs that would in turn result in more effective search. In CoRee, the context of the query is retrieved by traversing the already created UNL based indexer. The frequently occurring UNL relations obtained from the UNL index, in effect provide information about the possible connections between concepts in the specific domain under consideration. These connections provide the context of the query concept, and the query expansion based on this context yields meaningful search results. This context based query expansion is explained below.

7.2 Context based Query Expansion

In CoRee, the context of a query concept is defined as the association of this concept with other concepts in a CRC relation, across documents in the domain of interest. By analyzing the index, the concept associated with a query is matched with the CRCs of the index and the most common CRCs associated with the query concept are extracted. The expanded concepts obtained, are ranked based on frequency of CRC and on its being an entity. Query expansion is an on-line activity and the index analysis results in efficient query expansion. The most frequently occurring CRC in the index indicates the frequent association of concepts in the domain across documents and hence gives the domain context of the query concept. This expansion of the query concepts to CRC allows context dictated query sub graphs to be constructed for the query [46]. The expanded query graph is now associated with actual query terms, query concepts and expanded concepts associated with the context of the query concept.

The index based query expansion influences the searching and ranking of documents in many ways. The association of expanded concepts with the query, helps to build CRC query graphs that can be matched with the UNL index. Without this expansion, single word queries would have resulted in isolated concept (C) only match while with the expansion we are matching with a context dictated CRC.

8. Conceptual Searching and Ranking

The basic searching procedure is based on complete CRC Match or partial CR or C matches between query sub graphs and the

corresponding index as in AgroExplorer [8]. However, in CoRee, the design of the ranking procedure depends on whether the match of the index is with the actual query terms, actual query concepts or expanded concepts. In addition, all the sentence and document based features associated with the conceptual indices also affect the ranking procedure. Hence, the overall algorithm for ranking is a three level ranking.

The first level ranking is obtained based on whether there is complete match (CRC match), partial match of Concept Relation (CR) or match of only concepts (C Only). This level of ranking is provided by the Degree of Match Categorization tag Ta. Ta indicates the extent of match between the CRC representing the query sub graph and the conceptual index. It essentially differentiates between CRC, CR and C matches. The UNL sub graph is a directional graph and hence partial match also considers whether the concept in CR (Concept Relation), matches with the source concept, Cxi , or destination concept, Cyi , of the UNL subgraph. Let D:Ta be the set of matched documents associated with each Ta.

This set of documents obtained in level 1 category is further prioritized using Concept Association Categorization Tag Tb. Concept Association categorization depends on whether the index match is between query terms, query concepts or expanded concepts. In other words Tb is determined based on whether the Ci value in CRC, CR and C matches correspond to the actual query term, the concept of the query term, or the concept obtained after query expansion [46]. Accordingly the concept association is said to be of three types:

- Query Term TWi association - This means that the concept Ci is query term itself
- Concept Word CWi association - This means that the concept Ci matches the corresponding concept of the query, but the actual query term is different.
- Expanded Word EWi association - This means that the concept Ci is associated with a concept that is not actually in the query but has been obtained as a result of query expansion.

Once the documents have been ranked by Ta and Tb, the documents at the same Ta.Tb level are ranked based on weights calculated using the other features associated with the concept. The features used are position, frequency count, Named Entity (NE) tag and Multi-word (MW) tag of the term/concept. Thus the conceptual searching and ranking algorithm considers degree of match, context of query, concept association and index based term, concept and position factors corresponding to sentences as well as documents for effective ranking.

8.1 Evaluation of Search

CoRee has been tested with a corpus of 33000 documents from tourism domain We have used MAP (Mean Average Precision) (ThomJ et.al., 2007) for evaluation. The relevance judgment of each retrieved page is done based on the human judgment of the documents.

A query set of 139 queries has been used for the evaluation. The comparison of UNL search with that of key word based search based on precision at top 5, 10 and 20 documents is shown in Figure 6. It can be seen that the precision score of the UNL system is found to be around 0.72 even at the top 20 documents, while the key-word based search has a precision of 0.32. This shows the effectiveness of the UNL based Search .

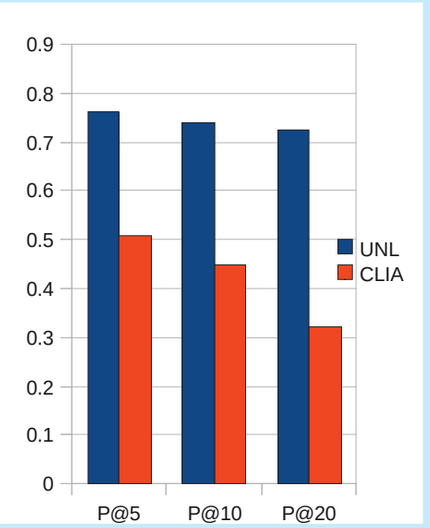


Figure 6: Comparison of UNL search with CLIA – word based search

9. Conclusion and Future work

The distinct features of the CoRee semantic search framework can be summarized as follows:

- Use of UNL based semantics in every component of the Search Engine – document representation, indexing, searching, ranking, query translation, query expansion and summary generation
- Use of morpho-semantic features for rule based approach in UNL enconversion
- Use of UNL based word semantics and sentence structure for anaphora resolution
- Design of three types of indices concept (C), Concept-relation (C-R) and (Concept-Relation-Concept (C-R-C) with sentence and concept specific features incorporated to aid in searching and ranking
- Use of conceptual index analysis for context based query expansion
- Design of a three level ranking technique based on degree of match, term-concept-expanded concept associations, and weight of index based features
- The design of a template based UNL oriented summarizer

Another significant outcome of this work is the building of language resources – namely – UNL list for tourism domain.

The experience gained from this effort has thrown open a number of challenging issues that need to be further addressed in UNL based search. One of the first issues to be addressed is the handling of nested UNL relations, co-reference resolution, and extensive word sense disambiguation for a richer semantic representation of the documents. In turn, this requires the design of Indices and a search mechanism to tackle nested UNL relations, which is a challenging task. Ranking is a task that requires constant innovation. In the context of CoRee, ranking can be further improved by differentiating UNL semantic constraints and by using associated UNL based ontology. Additionally statistical learning of categorization and priority of relations can be explored for use in ranking. The output processing can also be further improved by providing query based UNL oriented summary. Another important aspect to be addressed is the use of the UNL based framework for cross-lingual information access. We need to develop language dependent portions of the framework for other languages such as English, and evaluate the design of CoRee as a cross-lingual platform.

Acknowledgement

We thank the Ministry of Communications and Information Technology, DIT, New Delhi for funding this project under the consortium for the development of Cross-Lingual Information Access.

References

[1] Tumer, D. Shah, M.A. Bitirim, Y.Dept. of Comput. Eng., Eastern Mediterranean Univ., Famagusta -An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia", Internet Monitoring and Protection,. ICIMP '09. 2009 **page no:** 51- 55 **Location:** Venice/ Mestre,

[2] Brin, S. and Page, LThe anatomy of a large-scale hyper textual web search engine. In Computer Networks and ISDN Systems. 107117, 1998

[3] Nomura, S, Oyama, S, Hayamizu, T, Ishida, T,Dept. of Social Informatics, Kyoto Univ, Analysis and improvement of HITS algorithm for detecting Web communities, Applications and the Internet, 2002. (SAINT 2002). Proceedings. 2002 Symposium on 2002,Pages:132–140, ISBN:0-7695-1447-2

[4] Susan L. Price, Marianne Lykke Nielsen, Lois M. L. Delcambre, Peter Vedsted, Jeremy Steinhauer, Using semantic components to search for domain-specific documents: An evaluation from the system perspective and the user perspective,Journal Information Systems archive Volume 34 Issue 8, December, 2009.

[5] Uchida H., Zhu M., The Universal Networking Language (UNL) Specification Version 3.0 1998.Technical Report, UNU, Tokyo. 1998

[6] UNDL. 2009. Universal networking digital language. http: // www. undl. org/ Online; accessed 28 September 2009

[7] Virach Sornlertlamvanich, Tanapong Potipiti and Thatsanee Charoenporn, 2001 NECTEC, Thailand, UNL Document Summarization, The First International Workshop on MultiMedia Annotation, 30-31 January 2001, Tokyo, Japan

[8] Mrugank Surve, Satish Kagathara, Pushpak Bhattacharyya, Agro Explorer Group, Agro Explorer: a Meaning Based Multilingual Search Engine, In Proceedings of the International Conference on Digital Libraries (ICDL), Volume 2, New Delhi, India, 2004

[9] Donald Metzler, Beyond Bags of Words: Effectively Modeling Dependence and Features in Information Retrieval,SIGIR Forum,volume 42, 1st issue, issn 0163-5840, pages = 77--77,2008

[10] Evgeniy Gabrilovich, Shaul Markovitch, Wikipedia-based Semantic Interpretation for Natural Language Processing, Journal of Artificial Intelligence Research 34 2009 443-498

[11] HAKIA.(2009).A Hakia search engine. http://www.hakia.com Online;accessed 28 september 2009.

[12] Kraines, S., Guo, W., Kemper, B., And Naka-Mural, Y. EKOSS: A Knowledge-User Centered Approach to Knowledge Sharing, Discovery,and Integration on the Semantic Web. Springer Berlin / Heidelberg, 2006.

[13] S Ensebot. Sensebot search engine. http: //www. sensebot. net/ Online; accessed 28 September 2009 .

[14] Gao, M., Liu C., and Chen, F An Ontology search engine based on semantic analysis. International Conference On the Convergence of Knowledge, Culture, Language and Information Technologies, 2005.

[15] Jain, M. and Damani, O. P. English to UNL (interlingua) Enconversion. Indian Institute of Technology Bombay, India, 2008.

[16] Gala, N. Using an incremental robust parser to automatically generate semantic graph. Proceedings of the 3rd workshop on Robust Methods of Analysis of Natural Language Data, 2004.

[17] Igor M. Boguslavsky, Leonid L. Iomdin , Victor G. Sizov, Interactive enconversion by means of the ETAP-3 system, Proceedings of the International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies, 2003.

[18] Sameh Alansary, Magdy Nagi and Noha Adly, A Library Information System (LIS) based on UNL knowledge infrastructure, Seventh International Conference on Computer Science and Information Technologies, 28 September - 2 October, Yerevan, Armenia, 2009.

[19] RajatKumar Mohanty, Anupama Dutta, P. B. Semantically relatable sets: Building blocks for representing semantics. Machine Translation Summit, 2005.

[20] Nguyen, D.P.T. Ishizuka, M.A A Statistical approach for universal networking language-based relation extraction Research, Innovation and Vision for the Future, IEEE International Conference, 2006.

[21] Ali, M.N.Y. Al-Mamun, S.M.A. Das, J.K. Nurannabi, A.M., Dept. of CSE, East West_Univ., Dhaka. Morphological analysis of Bangla words for Universal Networking Language Digital Information Management. ICDIM 2008. Third International Conference, 2008.

[22] Firoz Mridha, Zakir Hossain, Shahid AI Noor, Development of Morphological Rules for Bangla Words for Universal Networking Language, IJCSNS International Journal of Computer Science and Network Security, vol. 10 No. 10, October 2010

[23] Nawab Yousuf Ali and Mohammad Zakir Hossain Sarker and Jugal Krishna Das, Analysis and Generation of Bengali Case Structure Constructs for Universal Networking Language, International Journal of Computer Applications, Vol. 18, March 2011, page no. 34-41

[24] Balaji J, Geetha T V, Ranjani Parthasarathi, Madhan Karky, Morpho-semantic features for Rule-based Tamil Enconversion, International Journal of Computer Applications, July 2011.

[25] S. E. Brennan, M. W. Friedman, C. J. Pollard.: A Centering Approach to Pronouns, Proceedings, 25th ACL, pp. 155-162, 1987.

[26] Hobbs, J. R.: Resolving Pronoun references, Lingua 44:311-338, 1978

[27] Grosz, B.J., A. K. Joshi, and S.Weinstein.: Centering: A framework for modeling the local coherence of discourse.

Computational Linguistics, 1995, page no. 202-225

[28] Shalom Lappin, Herbert J. Leass.: An Algorithm for Pronominal Anaphora Resolution, Association of Computational Linguistics. 1994.

[29] Sobha L.: Resolution of Pronominals in Tamil, Proceedings of the International Conference on Computing: Theory and Applications, ICCTA'07 ,IEEE, 2007.

[30] Narayana Murthi, K.N, Sobha, L, Muthukumari, B.: Pronominal Resolution in Tamil Using Machine Learning Approach, The First Workshop on Anaphora Resolution (WAR I),Ed Christer Johansson, Cambridge Scholars Publishing, 15 Angerton Gardens, Newcastle, NE5 2JA, UK, 2007, pp.39-50

[31] Tyne Liang and Dian-Song Wu.: Automatic Pronominal Anaphora Resolution in English Texts, Computational Linguistics and Chinese Language Processing, Vol. 9, No.1, February 2004, pp. 21-40

[32] Kamlesh Dutta, Nupur Prakash, and Saroj Kaushik.: Resolving Pronominal Anaphora in Hindi using Hobbs' Algorithm, Web Journal of Formal Computation and Cognitive Linguistics, Vol. 1, No. 10, January, 2008

[33] C.L. Yeh and Y.C. Chen.: An empirical study of zero anaphora resolution in chinese based on centering theory, Proc. ROCLING XIV, pp.1–18, Tainan, Taiwan , 2001

[34] Manuel Palomar and Antonio Ferrâ and Jes Âus and Peral Maximiliano Saiz-noeda and Rafael Muñoz.: An algorithm for anaphora resolution in Spanish texts, Association for Computational Linguistics Journal, vol- 27, 2001

[35] Richard Larson and Marta Luján.: Emphatic Pronouns, First distributed as "Focused Pronouns" MIT, 1984

[36] Pınar Tüfekçi and Yılmaz Kılıçaslan.: A Computational Model for Resolving Pronominal Anaphora in Turkish Using Hobbs' Naïve Algorithm, Proceedings of World Academy of Science, Engineering and Technology volume 5 april 2005

[37] KONG Fang, ZHOU GuoDong, ZHU Qiaoming.: Employing the Centering theory in Pronoun Resolution from the Semantic Perspective. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 987-996, Singapore 2009

[38] KONG Fang, ZHOU GuoDong, ZHU Qiaoming, QIAN Peide.: Using Semantic Roles for Corefernce Resolution, International Conference on Advanced Language Processing and Web Information Technology, IEEE 2008

[39] Balaji J, Geetha T V, Ranjani Parthasarathi, Madhan Karky, Anaphora Resolution Using Universal Networking Language, Indian International Conference on Artificial Intelligence, Dec- 2011

[40] Chen Bin, Su Jian, Tan Chew Lim.: A Twin-Candidate Based Approach for Event Pronoun Resolution using Composite Kernel, Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 188–196, Beijing, August 2010

[41] http://en.wikipedia.org

[42] Maxim Martynov, Boris Novikov, An Indexing Algo-rithm for Text Retrieval, Proceedings of the Third Interna-tional Workshop on Advances in Databases and Informa-tion Systems – ADBIS'96, 1996

[43] Mark Allen Weiss,Data Structures and Algorithm Analysis in C++ by Publisher: Addison Wesley Pub. Date: ISBN-13: 9780321441461, February 2006

[44] Rowena Chau,Chung-Hsing Yeh,Fuzzy Conceptual Indexing for Concept-Based Cross-Lingual Text Retrieval, IEEE Computer Society, September/October 2004 (Vol. 8, No. 5) ISSN: 1089-7801

[45] Subalalitha, T.V.Geetha, Parthasarathi, R., and Karky, M. CoReX: A Concept Based Semantic Indexing Technique. SWM-08, 2008

[46] E Umamaheswari, T.V.Geetha, Parthasarathi, R., and Karky, A Multilevel UNL Concept based searching and ranking- (WEBIST 2011)-(WEBIST) 7th International Conference on Web Information Systems and Technologies 2011