Cross-lingual information retrieval and delivery using community mobile networks

R. Shriram[1]
Vijayan Sugumaran[2]
Epaminondas Kapetanios[3]

[1] TIFAC-CORE on Pervasive Computing, Velammal Engineering College
[2] Decision and Info Sciences School of Business, Oakland University
[3] Harrow School of Computer Science, University of Westminster

Cross-Lingual Information Retrieval and Delivery Using Community Mobile Networks

R. Shriram
TIFAC-CORE on Pervasive Computing
Velammal Engineering College
Chennai – 600066, India
shrionsong@yahoo.com

Vijayan Sugumaran
Decision and Info Sciences
School of Business
Oakland University
Rochester, MI 48309
sugumara@oakland.edu

Epaminondas Kapetanios
School of Computer Science
University of Westminster
309 Regent Street
London W1B 2UW-UK
e.kapetanios@wmin.ac.uk

Abstract

Much of the Web content is in English and accessing this content is difficult for non-English speaking users because of the language barrier. Hence, there is a great need for providing applications and interfaces in one's own language to tap into this vast knowledge reserve. In addition, access to the Internet is still a major problem in developing countries because of the "digital divide" and hand held devices such as PDAs and Mobile Phones are seen as enablers in bridging this gap. However, displaying cross-lingual content on these mobile devices is a non trivial task and there is a great need for robust mechanisms and infrastructure for content delivery in different languages on the fly. This paper presents an overall approach for cross-lingual content specification and delivery for computing/mobile devices. It helps mitigate the language barrier by providing cross-lingual search and retrieval capabilities for accessing the Web content.

1. Introduction

The World Wide Web has evolved into a tremendous source of knowledge and it continues to grow at an exponential rate. According to [11], 68% of Web content is in English. While a vast amount of information is available on the Internet, much of it is inaccessible because of language barrier. In order to make this knowledge resource available to non-English speaking segment, search applications that use native language interfaces are needed. Users should be able to specify search queries in their own language in order to retrieve documents and useful information from the Internet. This paper lays the foundation for such an application with the development of an approach for cross-lingual web querying and summarization.

In order to provide cross-lingual content to end users, a number of research questions have to be answered. For example, how do we provide content in different languages from a single information source? How do we make the approach flexible and scalable so that multiple languages could be accommodated? How do we interface with existing commercial search engines such as Google and summarize the relevant information sought by the user in his or her language? Good solutions to these problems are needed in order to tap into the vast Internet resource. We envisage the following scenario. The search query in a specific language is parsed and disambiguated using the lexicons available in that language and a query tree is constructed. This query tree can then cloned into any target language and submitted to a search engine. If no language resources are available, the initial search terms are translated into English. These English terms may be further disambiguated using WordNet and other ontologies and the expanded query is submitted to the search engine. The results from the search engine are summarized using a meta-language. It is then mapped to the target language and the results presented to the user. We contend that one of the major advantages of our approach is scalability and hence the methodology can be expanded to other languages.

The specific objectives of this research are:
i) developing a methodology for cross-lingual Web querying and retrieval and automatically summarizing the content from Web resources,
ii) developing a content specification meta-language that can be used to represent search result content and mapping it to a target language.

1-4244-0682-X/06/$20.00 ©2006 IEEE

The contributions of this research are four fold. First, the content specification meta-language facilitates capturing content in only one format as opposed to different languages. Second, the porting mechanism clearly lays out the translation mechanism that would be used to dynamically translate the content into the target language on the fly. Third, the cross-lingual web querying methodology will facilitate easy access to Web content and minimize the language barrier. Fourth, the resulting infrastructure design to create a community mobile network can be applied to other application domains.

2. Literature Review

Search by important associations to some related concepts or user context or profile has been addressed by the emerging concept or knowledge based querying approaches [12, 13]. In general, these approaches make use of domain specific ontologies and semantic annotations in order to augment and improve query semantics either interactively [3] or as performed by the system [12, 13] for different purposes, e.g., search engines and information retrieval [13], and mediation across heterogeneous data sources [12]. However, they mostly rely on intelligent techniques and knowledge-based approaches for mappings across concepts and query expansion. Querying by integrating semantic associations among entities, instances, properties, etc., into a conceptual search or query language has not been addressed, especially when cross-lingual natural language based querying is concerned.

Within the realm of cross-lingual information retrieval (CLIR) [14, 15], a number of challenges have been reported, especially the problem of query translation [15]. To make query translation possible, existing IR systems rely on bilingual dictionaries for cross-lingual retrieval. In these systems, queries submitted in a source language are translated into a target language using simple dictionary lookup. If this is not possible, query translation is performed by corpus-based techniques [14] in which translation equivalents are extracted from parallel corpuses.

Research on technology development for Indian languages is on the rise [8, 9] and there is great interest in developing tools for computing in various Indian languages. A few CLIR efforts have been reported in the literature, particularly for Hindi [6, 7]. Larkey et al. [6] describe an approach that uses two different probabilistic retrieval models. Their approach uses simple translation of the query terms without any refinement. Xu and Weischedel [7] also provide a probabilistic model based approach for cross-lingual retrieval for Hindi. Both of these efforts report only limited success because of ambiguity and loss of information during syntactic based translation.

Kumaran [2] has proposed a new flexible architecture – Multilingual Information processing on Relational Architecture (MIRA) – that supports the multilingual processing functionality of the primary storage mechanism, namely, the relational database systems. While this work has some similarity with our current research project, his approach is narrow in the sense that it handles only content stored in a relational database where excellent querying facilities already exists. Our work is targeted towards retrieving information from the Web, which is heterogeneous.

Kagathara et al. [4] describe a special purpose search engine for the Agricultural domain called AgroExplorer, which is designed to search and retrieve the contextual information relevant to the users in their own languages. In order to facilitate this functionality, the system extracts the meaning of a query which is represented in the form of Universal Networking Language (UNL) expressions. Our work is not domain specific and hence can be applied to any domain.

Devi et al [5] discuss the details of a Tamil Search Engine and discuss the issues related to the crawler, database storage architecture and other functional modules of the search engine. While they have shown some limited success, the approach used by their search engine is limited to the Tamil language. Our work is capable of handling multiple languages.

Kumar et al., [22] propose PICO, a framework for creating mission-oriented dynamic communities of autonomous software entities that perform tasks for users and devices. Our content delivery approach shares some similarities such as the community network, handling dynamic information and selective content delivery. However we use messaging mechanisms based on standard infrastructure and our application is unique in that it is information intensive.

3. Proposed Methodology

The solution we envision contains the following two major elements: a) cross-lingual information retrieval methodology, b) content meta-language and mapping to target language. Each of these elements and the resources that are needed to develop the overall solution are described below.

3.1 Cross-Lingual Information Retrieval

This research proposes a Cross-Lingual Information Retrieval approach that is used to search

Internet resources for appropriate content and summarize it in a succinct form using the content specification meta-language, developed as part of this research. This content is then mapped to the target language. We focus on developing a methodology for querying the Web in languages other than English, namely Tamil to start with, and retrieve relevant documents, translate and summarize them and present the information to the user in Tamil. Several efforts have been undertaken with respect to using Tamil for social computing on the Internet [5, 10]. Our research builds on the results in the areas of syntactic parsing of Tamil [10] and Tamil search engine [5] in developing our CLIR system for Tamil and other south Indian languages that interface with search engines such as Google.

Our concept based cross-lingual web querying adapts the semantic query augmentation method discussed in [1]. It consists of the following five steps: a) MDDQL Query Parsing, b) Query Expansion, c) Query Formulation, d) Search Knowledge Sources, and e) Translate and Summarize Results. The system architecture that implements our cross-lingual web querying is shown in Figure 1. The individual steps of our methodology are briefly described below.

Step 1 - MDDQL Query Parsing: The first step involves parsing the natural language query specified by the user. The query is segmented using an appropriate segmentation tool that is available for the language used to specify the query. The MDDQL parser parses the segments and creates the query graph. The parsing algorithm does not assume any underlying grammar, but adjusts its behavior based on the parameters specified. These parameters represent the hidden rules of all grammars.

The importance of this design decision is that it allows the system to dynamically vary the language it is tokenizing without changing the parser itself. The vertices in the conceptual query graph created are eventually translated into English (or any other target language) to generate the initial query terms. Note that the initial query is not translated into English immediately. The intention is to postpone it as much as possible, since all translations are prone to ambiguities if the context is not clearly specified and the words are not chosen carefully from a lexicon. For example, if the query is written in Tamil, the words are disambiguated using Tamil WordNet [21] and other existing ontologies before translating them into English. A complete description of our MDDQL parsing algorithm and its implementation is provided in [16].

Step 2 – Query Expansion: The output of the MDDQL parsing step is a set of initial query terms which become the input to the query expansion step. The query expansion process involves expanding the initial query using lexicons and ontologies. It also includes adding appropriate personal information as well as contextual information. For each query term, the first task is to identify the proper semantics of the term, given the user's context. To do so, the word senses from a lexicon (WordNet for English, TamilWordNet for Tamil, etc.) are used. For each term, synonym sets are extracted. The appropriate word sense is

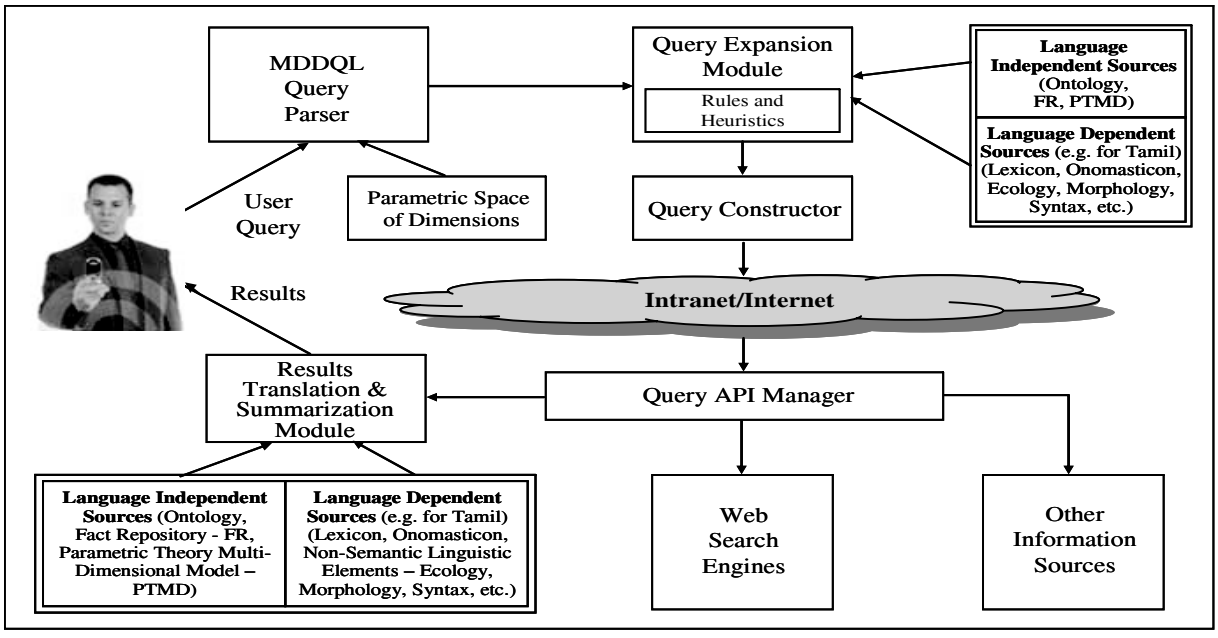

Figure 1. System Architecture for Cross-Lingual Web Querying

determined based on the context and other query terms (may also need user input) and a synonym from that synset is added to the query. To ensure precise query results, it is important to filter out pages that contain incorrect senses of each term. Thus, a synonym from the unselected synset with the highest frequency is added as negative knowledge to the query.

Step 3 –Query Formulation The output of the query expansion step is the expanded set of query terms that includes the initial query terms, synonyms, negative knowledge, hypernyms, hyponyms, and personal preference information. This expanded set becomes the input to the query formulation phase. In this phase, the query is formulated according to the syntax of the search engine used. Appropriate boolean operators are used to construct the query depending upon the type of term added. For each query term, the synonym is added with an OR operator (e.g. query term OR synonym). Hypernym and hyponym are added using the AND operator (e.g. query term AND (hypernym OR hypernym)). Personal preference is also added using the AND operator (e.g. query term AND preference). The negative knowledge is added using the NOT operator. The first synonym from the highest remaining synset not selected is included with the NOT operator (e.g. query term NOT synonym).

Step 4 – Search Knowledge Sources This step submits the query to one or more web search engines (in their required syntax) for processing using the API provided by them. Our query construction heuristics work with most search engines. For example, AltaVista allows queries to use a NEAR constraint, but since other search engines such as Google and AlltheWeb do not, it is not used. Likewise, query expansion techniques in traditional information retrieval systems can add up to 800 terms to the query with varying weights. This is not used in our approach since search engines limit the number of query terms.

Step 5 – Translate and Summarize Results: In the final step, the results from the search engine (URLs and 'snippets' provided from the web pages) are retrieved, translated using the meta-language and mappings to the target language. Available lexicons and ontologies are also used in the translation. The summarized content is then presented to the user. The user can either accept the results or rewrite and resubmit the query to get more relevant results.

4. Role of Content Meta-language for Query and Search Result Translation

Having presented the major phases of the concept based cross-lingual information retrieval approach, we embark on illustrating the underlying philosophy for the translation of queries or search result from a Natural Language Processing (NLP) point of view.

Figure 2 depicts a reference architecture for NLP, which indicates that translation should take place in terms of extracting and representing text meaning rather than simply using dictionaries. To this extent, the semantic equivalence of the translated phrase or text to the original content is improved.

Text meaning representation is usually performed in terms of a meta-language as a constrained set of linguistic patterns and structures which are mapped to target languages by applying appropriate algorithms and heuristics. Text meaning representation (TMR) has always been a key issue in traditional NLP system architectures such as the stratified model or the flat one. All these architectures aimed at extracting of a TMR from some input text or generate output text from it in a modular way. This is achieved either by running the modules, e.g., ecological analysis, morphological analysis, syntactic analysis, lexical semantic analysis, discourse/pragmatic analysis, on a text or TMR one by one (stratified model) [17], or simultaneously without waiting the results of previous modules (flat model) [18]. Constraint-Satisfaction NLP architectures [19] allow the exploitation of module specific results in the 'flat' model as posed constraints for other modules. Figure 2 depicts a modularized, not pipelined architecture, e.g., flat and *constraint satisfaction* NLP architectures for the extraction of text meaning.

The complexity and expressiveness of the chosen TMR structure depends on the input text at hand. Ideally, the addressed TMR meta-language needs to reflect and capture both kinds of meaning in natural language: static and dynamic [20]. The former resides in lexical units (morphemes, words, phrasals). The latter resides in meaning of clauses, sentences, paragraphs and larger text units. In theory, TMR should provide the specification of how, for a given text, static, context independent meanings of its elements are combined into a dynamic, context dependent text meaning representation and vice-versa.

For the NLP reference architecture (Figure 2), the natural language based query is a type of input text. Parsing of this text across all aspects of NLP, from ecology to pragmatics, aims at generating the query meaning representation. This is achieved in terms of a high level, conceptual query tree, which enables the capture of ontological semantics of the query terms. Consequently, query translation takes place in terms of generating queries in a specific natural language from the query meaning representation rather than following a word-by-word translation of the query.

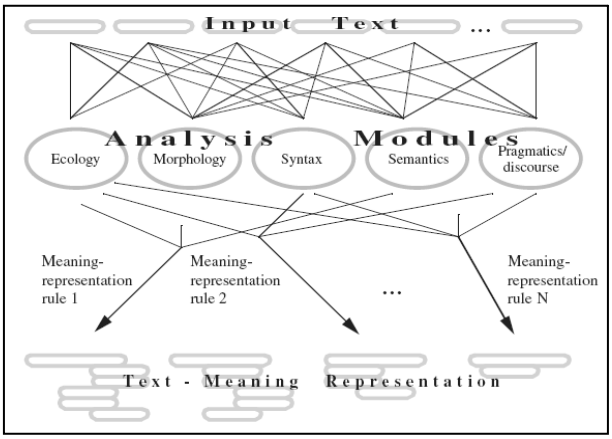

Figure 2. A Modularized Architecture for Text Meaning

5. Community Mobile Network for Content Delivery

One of the methods to deliver multilingual content is via the community network. The architecture of the community network is shown in Figure 3. The community network is equipped with specific *mobile device application* software at the user's end. The interfaces in the software allow the language to be selected with proviso for minority languages not in the phone's hardware. As the customized interface is uniform across mobile devices, the user can also access personalized information and content. In addition, the application must allow mobility across languages.

One key aspect in the content provider services is the ability to handle different types of events when they occur. For this, the content provider service database is maintained. It allows the clients to locate and access services and applications seamlessly. The database also stores a list of user preferences (rules) that are to be consulted when an event occurs.

The communication components in the network infrastructure in conjunction with the application software in the mobile device allow seamless information querying and delivery. The applications in the mobile device encode the information and enable transmission/reception of information in a language independent manner. The architecture works on the request/response principle based on messaging gateways. This allows the community network to leverage traditional messaging infrastructure instead of the user needing internet access on the mobile devices.

The defining aspect of mobile computer users is that they can be in different contexts, which we may model as ***active data space.*** Users may wish different behaviors from their mobile devices depending upon linguistic preferences, location and behavior context. Also the requests for static information must be treated differently from a request for dynamic or real time information. Context is provided, both synchronously and asynchronously, as context events. Context components, in turn, utilize the available static and dynamic preferences for acquiring contextual data from the environment. The active spaces corresponding to each user is maintained on the server that tracks not only the personalized context aware information access, but also acts as a lower level cache for other users in case of non-secure general data. The personal profiles at the device and the mobile network level are synchronized at regular intervals to keep pace with the changing preferences of the users.

The Internet interface provides a framework for information retrieval. This internet interface works in tandem with the intelligent reasoning mechanisms for accessing real time information like cricket scores or stock market quotes. The other aspect of the intelligent reasoning mechanism which works in tandem with the CLIR and the Text meaning systems is that it allows translation of information on the fly.

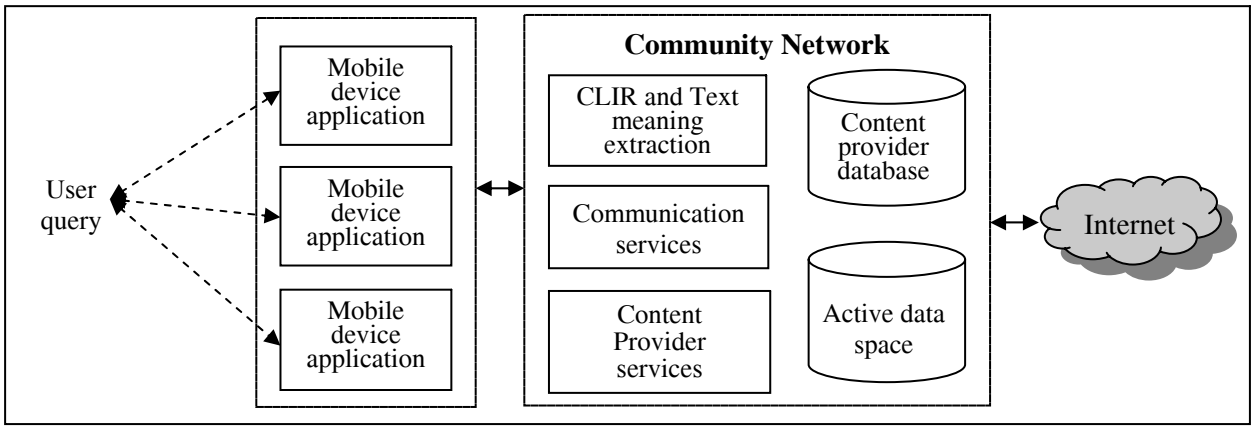

Figure 3. Community network architecture

6. Conclusion

While a few retrieval techniques exist for searching in different languages, they are not adequate for performing multilingual querying on the Web because of their inability to handle the heterogeneity of the format and structure of Web documents. In this paper, we have presented a concept based methodology for multilingual querying on the Web. Our MDDQL-based approach is language independent and hence the approach is scaleable. It also uses contextual and semantic information in query refinement to improve precision. We have discussed the architecture of a system that implements our methodology. A proof-of-concept prototype is currently under development to demonstrate its feasibility. Our future work includes completing the prototype, experimental validation of the prototype, and further refinement of the methodology. Advances in multilingual Web querying will move us a step closer to making the Semantic Web a reality.

7. Acknowledgement

We wish to acknowledge with thanks the TIFAC-CORE on Pervasive Computing for the infrastructural support in this project.

8. References

[1] Burton-Jones, A., Storey, V., Sugumaran, V., Purao, S. "A Heuristic-based Methodology for Semantic Augmentation of User Queries on the Web," *22nd International Conference on Conceptual Modeling*, Chicago, Illinois, October 13 – 16, 2003, pp. 476 – 489.

[2] Kumaran, A. "MIRA: Multilingual Information Processing on Relational Architecture," Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India.

[3] Fonseca.B, Golgher.P, Possas.B, Ribeiro-Neto.B, Ziviani.N, Concept-based Interactive Query Expansion, *CIKM 2005*, pp. 696-703.

[4] Satish Kagathara, Manish Deolalkar, Pushpak Bhattacharyya, "A Multi Stage Fall-back Search Strategy for Cross-Lingual Information Retrieval," *Working Paper*, Media Lab Asia, KReSIT, IIT Bombay, http://www.mlasia.iitb.ac.in/docs/SIMPLE-IITB-multi-stage-search-camera_ready-27dec04.pdf

[5] Deepa Devi.J, Parthasarathi.R, and Geetha T.V, "Tamil Search Engine," *Tamil Internet 2003*, Chennai, Tamilnadu, India.

[6] Larkey, L.S., Connell, M.E., Abdul Jaleel, N. "Hindi CLIR in Thirty Days," *ACM Transactions on Asian Language Information Processing*, Vol. 2, No. 2, 2003, pp. 130 – 142.

[7] Xu, J., and Weischeel, R. "Cross-Lingual Retrieval for Hindi," *ACM Trans on Asian Language Info Proc*, Vol. 2, No. 1, 2003, pp. 164 – 168.

[8] Bharati, A., Chaitanya, V., Sangal, R. "Computational Linguistics in India: An Overview," http://acl.ldc.upenn.edu//P/P00/P00-1077.pdf

[9] Vikas, O. "Language Technology Development in India,"http://www.indictrans.org/Articles/English/article_src/Indic/ncst2.pdf

[10] Saravanan, K., Parthasarathi, R., Geetha, T.V. "Syntactic Parser for Tamil," *Proceedings of the Tamil Internet Conference, Chennai*, Tamilnadu, India, August 22 – 24, 2003, pp. 28 – 37.

[11] Global Reach: "Global Internet Statistics," (2005) http://global-reach.biz/globstats/index.php3

[12] Sattler, K., Geist, I., Schallehn, E. Concept-based querying in mediator systems, Vol. 14, pp. 97–111, *VLDB Journal*, 2005.

[13] Liu, Z, Chu. W.W, "Knowledge Based Query Expansion to Support Scenario Specific Retrieval of Medical Free Text," *ACM SAC*, 2005, pp. 13-17.

[14] Wang, .J, Teng.J, Cheng.P, Lu.W, and Chien.L, Translating Unknown Cross-Lingual Queries in Digital Libraries Using a Web-based Approach, *JCDL'04*, 2004, pp. 108-116.

[15] Lu, W, Chien.L, Lee.H, Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach, *ACM Trans. on Inf. Sys*., Vol. 22, No. 2, 2004, pp. 242-269.

[16] Kapetanios, E, Sugumaran, V., Tanase, D. "Multi-Lingual Web Querying: A Parametric Linguistics Based Approach," *NLDB2006*, Klagenfurt, Austria, May 31 – June 2, pp. 94 – 105.

[17] Roger, C. S., Eugene Charniak, Yorick Wilks, Terry Winograd, William A. Woods, Natural Language Processing. *IJCAI 1977*: 1007-1013.

[18] Ballim, A., Wilks, Y., Barnden, J.A. Belief Ascription, Metaphor, and Intensional Identification. *Cognitive Science* 15(1): 133-171 (1991).

[19] Nirenburg, S., Frederking, R. E., Farwell, D., Wilks, Y. "Two Types of Adaptive MT Environments," *COLING* 1994: 125-128.

[20] Nirenburg, S., Raskin, V. Ontological Semantics. *MIT Press*, Sept. 2004, ISBN 0-262-14086-1

[21] Thiyagarajan, .S, Arulmozi., S, Rajendran, .S, "Tamil WordNet," *First Global WordNet Conference, CIIL*, Mysore, 21-25 Jan 2002.

[22] Kumar, M., Shirazi, B.A., Das, S.K., Sung, B.Y., Levine, D., Singhal, M. "PICO: A Middleware Framework for Pervasive Computing", *IEEE Pervasive Computing*, pp 72-79, July-Sept. 2003