Input Processing for Cross Language Information Access

Vasudeva Varma, Bhupal Reddy, Aditya Mogadala, Srikanth Reddy Vaddepally,
Nikhil Priyatam Pattisapu, Mahathi Bhagavatula
Search and Information Extraction Lab
International Institute of Information Technology (IIIT)
Hyderabad, India
vv@iiit.ac.in, {bhupal_iiit, aditya.mogadala, srikanthreddy.v, nikhil.priyatam,
mahathi.b}@research.iiit.ac.in

Abstract:- ***India is the home for many languages each having its own rich set of syntax and semantic rules. With the increasing popularity of the Internet, the amount of Indian language content on the web is also growing. Indian language publishing portals used custom techniques and proprietary fonts to present their content before Unicode was popular. Thus an abundance of information has been trapped in proprietary encodings and partitioned across multiple languages. Thus, it limited the amount of information a person can access to the content available in language(s) they know. Cross Lingual Information Access (CLIA) systems need to address these issues.***

In a CLIA system, the major focus is on presenting information to the user in a language they know along with fetching documents from multiple languages. Building such a system has common problems like 1) Identification of a web page language so it can be processed with a language specific analyzer, 2) Forming the right queries in the target language given that the language of the query and the language of the documents to be retrieved are different in a CLIR system and 3) Translating the snippets, summaries of web pages into the users language.

In this paper we discuss the techniques that have been developed as part of the CLIA project for solving a few of these problems for Indian languages.

Language Identification; Domain Identification; Input Processing; Query translation; Query Transliteration; Query-Formation

I. Introduction

Information Access is the process of making the information available in various documents accessible and usable to the user who have a specific information need. The documents may be of various media, formats, document sources or even languages. Information Retrieval (IR) technologies enable information access by retrieving a set of ranked documents that are likely to be relevant to the information need of the user. However, IR is only a part of the information access puzzle. Role of IR technologies ends once the relevant documents are obtained. After the results are obtained, the user needs to skim through the documents, judge the relevance of these documents, compare them against each other, find out relevant portions of the document that might satisfy their information need, extract elements of the text that provide answers, and perhaps summarize multiple documents or portions of the documents. All this requires processing of world knowledge. Information Access technologies are expected to provide these functionalities that is more cognitive in nature.

CLIR can be seen also as a technology

that combines both Information Retrieval and Machine Translation (MT) and extend to the languages other than English. The structure of CLIR system is broken down into categories like Indexing (IR), translation, ranking and matching. Oard & Dorr [1] work is perhaps the first attempt to study various approaches and techniques of CLIR in a detailed manner and showed that CLIR is not exactly the combination of IR and MT and somewhere between them. IR focuses on retrieving the relevant documents given a query by a human user and MT aims at producing single accurate target equivalent of a given input text. CLIR's functionality may be a hybrid of both these systems in the sense that the IR engine of CLIR system can take multiple translations produced by a program (as opposed to human user's single query) and the translation system tries to process not so complete input text (for example, just named entities, multiword expressions or simple sentence fragments) and produce multiple target language equivalents focusing less on producing grammatically correct translations.

In this paper we mainly focus on three important aspects or modules of CLIA system (a) input query processing (b) documents crawling by taking language identification and domain identification into consideration and (c) query formation. We divide the above-mentioned areas into different sections. First we describe the overall challenges in building the CLIA systems in Section II. Then we see some preprocessing aspects like language identification in Section III and domain identification in Section IV during crawling. Query processing aspects like query translation, transliteration, formation and disambiguation issues are covered in Section V. In Section VI we explain the experimental results obtained in each of the modules mentioned. Section VII covers the conclusion and future aspirations.

II. Architecture of the CLIA system

The functionality of the CLIA system can be broken down into two main phases – A) Offline Processing and B) Online Processing.

A. Offline Processing

The offline-processing phase encapsulates the following tasks – crawling, parsing and indexing.

1) Crawling

This task simply refers to the process of crawling the web and fetching all the documents found for further processing. The crawl starts by taking as input a set of starting URLs called as Seed URLs. Once these URLs are fetched, the out links from these documents are added to the queue and the crawler fetches them too. This process is repeated until the queue is empty or a present limit is reached.

2) Parsing

The parsing phase is an important part of the offline phase. All processing of the fetched documents is done in this phase. The CLIA system does the following tasks during this phase not necessarily in the specified order.

- Font Transcoding: Conversion of web pages with non-standard or proprietary character or font encoding into standard UTF-8.
- Language Identification: Identify the writing language of the document. Any language specific offline processing is done based on the language attribute of the document.
- Domain Identification: Categorize the webpage into one of the pre-decided domains. Current system supports only Tourism. This information is used for domain specific processing like deciding upon the IE template to be filled.

- Named Entity and Multi-word Expression Extraction: NEs and MWEs in the document are marked so as to facilitate other modules to utilize this information.
- Summary Generation: An extractive summary of the document is generated to benefit the user.
- Information Extraction: Domain specific IE templates are populated in this module for providing information to the user in a tabulated form.

3) Indexing

In the indexing phase, the various fields of a document are indexed and stored as an inverted index readily searchable. The important fields that are included in the index are the document title, content, language, domain and named entities. Document content is analyzed which includes stop word removal, stemming etc before indexing it.

The most important modules in the offline processing phase are the *Language Identification* module and the *Domain Identification* module. These are discussed in detail in sections III and IV.

B. Online Processing

Online processing refers to all the tasks that are done from the time the user has fired a search query to the time the user sees the results page. A brief note on all the online processing tasks of CLIA is provided below.

- Named Entity and Multi-Word Expression extraction: A light weight NER and MWE analyzer is used to extract the named entities and multi word expressions so that they will be directly searched in the index without any further analysis like stemming.
- Query Processing: For a monolingual IR system this includes query reformulation to improve recall. For a cross lingual IR system this includes conversion of query from one language into another.
- Retrieval and Ranking: Retrieve relevant documents from the index and rank them as per their importance.
- Snippet generation: Generating a small excerpt from the document highlighting the query context so that the user can take a decision whether to visit the webpage or not.
- Snippet translation: This is an important feature of the CLIA system wherein the snippet is translated into the user's language for easier understanding.
- Query focused summary generation: Generating summary of the chosen document focusing on the query context.

The entire functionality and performance of the CLIA system is dependant on the Query Processing module. This is discussed in detail in Section V.

III. Language Identification

Language Identification as the name itself suggests is the identification of the writing language of a web page. In the case of pages that contain multiple languages, the dominant language is taken as the language of that page. This is a crucial part for a Cross Lingual system as all other modules use language specific processing.

One of the major problems with Indian language web pages is that they use proprietary non-Unicode fonts or non-standard character encoding. The reason for this is that web pages in Indian languages existed even before the Unicode became a standard. To maintain consistency throughout the system, these non-standard pages are converted to Unicode during a preprocessing step while fetching. An open source tool called Unigateway (a PHP port of

Padma Converter) is used to do the transcoding. Once the conversion is done, all the other modules of the system, including the language identifier, work on the Unicode text. The other problem in identifying the language of the document is because of the wrong information in the metadata of the page this because of the common script they use across the domain. The problem in identifying language of the multi-language document is the hardest of all.

A hybrid technique has been implemented as part of the system to identify the language of the webpage. The system first looks for any available clues in the URL of the web page. These clues include language specific subdomain, (e.g. http://te.wikipedia.org), Indian language tokens in the URI etc. When the URL doesn't have sufficient clues about the language of the web page, the content will be processed. To identify language from content, N-gram profiles are used. The N-gram profile of a particular language contains the frequent n-gram words of that language. These files are generated by collecting the n-grams from large number of language specific documents in a separate process. The language of the webpage is identified as the language whose n-grams are present most in that page.

Character level information can also be used. However many Indian languages share a common character set but has different grammatical rules (e.g. Hindi and Marathi). The technique is still used as a filtering mechanism to reduce computation intensity by reducing the search space.

The CLIA system currently supports 7 languages viz. English, Hindi, Telugu, Tamil, Marathi, Bengali and Punjabi. Three new languages Oriya, Guajarati and Assamese are being incorporated into the system.

IV. Domain Identification

Domain Identification refers to the process of categorizing a webpage into one of the pre-decided domains. The current system has the ability to identify tourism related web pages. Health domain is being incorporated into the system. This information is mainly used to extract domain specific information from the web page and at a later stage generating domain specific rich snippets and summaries.

Processing URLs and content differently helps in domain identification. The system looks for certain domain specific words in the URL and returns a normalized score for that URL. Then the content of the web page is classified into one of the domains using a trained classifier. The current system uses an SVM classifier trained over tourism specific words from all the supported languages. The classifier returns a probabilistic score for that domain. A linear combination of the classification score and the confidence score from the URL is used to get the final score for the webpage. If it is greater than a particular threshold then the page is classified into the respective domain.

V. Query Processing

Query Processing is the most inevitable part of the CLIA system. After all, the overall cross-lingual performance of the system totally depends on how well the input query is in one language is converted to an equivalent query in some other language. The current CLIA system supports retrieval of Hindi and English documents, given a query in any of the supported Indian languages.

Query Processing consists of many trivial and non-trivial sub tasks. Trivial tasks include preprocessing steps like stop word removal, stemming, spelling correction etc. Stemming is not done on the source query because it affects

the translation and transliteration accuracies. However, stemming has to be done eventually once the query has been converted to the target language. Other trivial tasks include, query expansion and refinement, which have to be done both on the source query and the target query.

Some of the non-trivial tasks which are specific to the CLIA system are the – A) Query Translation, B) Query Transliteration, C) Query Formation and D) Query Disambiguation. Each of these tasks is discussed in detail below.

A. Query Translation

Translation is the straightforward technique of finding the equivalent target language word for a source query word. Query words that are other than named entities (NE) and multi-word expressions (MWE) needs to be translated from Indian language to one of the target languages i.e. Hindi or English. In order to translate them to cross language there is need of bilingual synset dictionaries in each of the Indian languages to English and Hindi. Machine Translation cannot be used directly here since MT is slow and the amount of information present in a query is not sufficient for getting an accurate automated translation. However, MT can be used for building bilingual synset dictionaries, which can then be used by the system for getting the translation.

There are bilingual synset dictionaries stored for each of the Indian languages mentioned earlier consisting at least 20k parallel words for translation. When the query is fired in an Indian language the words other than NE's and MWE's will be searched through the bilingual synset dictionaries for exact translation. Words for which the system couldn't find a translation, usually out of vocabulary words, are passed to the transliteration module before forming the final query.

B. Query Transliteration

Transliteration technique is applied to words that don't have a direct translation in the target language. Transliteration, the pronunciation-based translation or keyword-based translation from a source language to a desired target language, is important to many cross language and natural language processing tasks. Also success of any CLIA system depends on efficiency of its transliteration of queries. When not carefully handled, the mean average precision (MAP) can reach 50% [2]. Generally these queries may contain MWE's, NE's or general language terms. The challenge for the system is to identify the type of query terms and decide accordingly whether to translate or transliterate. Also, mining appropriate transliterations from the top results of the first-pass retrieval should be considered as it achieves enhanced cross lingual performance of the system overall, in addition to enhancing individual performance of more queries [3].

There are some drawbacks that need to be handled if transliteration does not produce the exact spelling variants used in the document collection. An approximate string matching technique is usually required to alleviate this drawback. By doing so, retrieval performance should not be compromised. If inaccurate or confusing transliterations occur in multiple alternative transliterations [4] appropriate action should be taken. Also, if translated or transliterated query is human readable it ensures proper transliteration and ultimately provides good search results.

For transliterating NE's in the first step we took word-aligned bilingual corpus while for the second step statistics over the alignments to transliterate the source language word and generate the desired number of target language words is done. For this Hidden Markov Model (HMM) alignment and Conditional Random

Fields (CRFs), a discriminative model is used. HMM alignment maximizes the probability of the observed word pairs using the expectation maximization algorithm and then the character level alignments (n-gram) are set to maximum posterior predictions of the model while CRF uses forward Viterbi and backward A* search whose combination produce the exact n-best results. For HMM alignment GIZA++ is used while CRF++ for training the model. The list of named entity words identified that need to be transliterated is taken and using the trained model top 3 probable words is extracted.

An experimental system is being developed which uses Romanization (plain character mappings across different languages) and nearest word match using Edit distance measure to predict the transliteration in English. This transliteration is then used to look up a parallel dictionary across all languages and get the equivalent transliteration in any of the supported languages. This technique, we believe, will not only alleviate the accuracy of the transliteration system but also expand the functionality of the CLIA system by providing the ability to transliterate words from any supported language to any other supported language.

C. Query Formation

Getting the translation or transliteration of all the words in the query is only part of the story. Often we end up with multiple translations or some fixed number of top transliterations for a query word. A proper query has to be formed from all the combinations generated, which are most similar, or equivalent to the source query in the target language has to be formulated. The Query Formation module is built to do this exact task.

The Query Formation module uses multiple techniques to generate a ranked list of equivalent queries. For this query disambiguation techniques and query performance prediction techniques are used. This module simply acts as a filter to reduce the amount of noise in the target language query. The better the translation and transliteration modules the better the formulated queries will be.

D. Query Disambiguation

Basic research is needed to investigate the relationship between sense ambiguity, disambiguation, and information retrieval. Understanding the influence of ambiguity and disambiguation on a probabilistic IR system is required to improve the retrieval results. Query ambiguity is only problematic to an IR system when it is retrieving from very short queries [5]. Also word senses in a query has to resolve to high degree of accuracy.

For CLIA, this creates much more complex problems as it involves understanding the query in multiple languages. To determine the sense of a word, a word sense disambiguation (WSD) algorithm typically should understand the context of the ambiguous word, external resources such as machine-readable dictionaries, or a combination of both. Although dictionaries provide useful word sense information and thesaurus provide additional information about relationships between words, they lack pragmatic information as unavailability of them in corpora. But there are major barriers in building a high-performing query disambiguation system as it includes difficulty of labeling data and predicting fine-grained sense distinctions.

User queries can be diversified into different topics and areas and could mean differently in each of those areas. This creates problem of word disambiguation. Thus any system that attempts to retrieve good results has to determine the sense of a word from contextual features.

VI. Experimental Results

The following sub sections discuss the evaluation techniques and experimental results of each of the modules discussed in the previous sections.

A. Language Identification

The language identification modules provide test code to evaluate its performance independent of the CLIA system. The performance of this module has been evaluated by giving plain monolingual documents as well as webpage's with mixed data. In almost all the cases, the language has been identified correctly. However, in exceptional cases like pages with equal amounts of content in multiple languages and pages with very little Indian language content, the module failed to determine the language. This drawback has been taken care of by analyzing clues from the URL of the webpage. Using this hybrid technique (discussed in Section III), the module was able to achieve an overall accuracy of 99% with a 1% loss due to erroneous or exceptional web pages.

B. Domain Identification

The domain identifier module currently can recognize only Tourism related documents. The module has been tested by giving a bunch of web pages which contained both tourism and non-tourism (noise) pages. The overall accuracy has bee calculated by analyzing how many documents was the module able to classify correctly. The classification accuracies of the module for all the languages are listed below.

TABLE I. DOMAIN IDENTIFICATION ACCURACY

Language	Accuracy
Telugu	83.5%
Hindi	66%
Tamil	90%
Marathi	95%
Bengali	63.2%
Punjabi	94%
English	68.5%

C. Query Processing

Cross lingual query processing performance of the CLIA system depends on the individual performance of the translation and the transliteration modules. Individual testing results of these modules are provided below.

1) Translation of Query

Since the translation module does a parallel synset lookup to translate words, the overall performance of this module depends on the quality and coverage of the dictionaries. Currently the system has parallel dictionaries containing about 50000 words each from all the supported languages to Hindi and English. The dictionaries include frequent words in each language and many tourism specific words.

2) Transliteration of Query

The transliteration module has been tested using the test data from *Named Entity Workshop (NEWS) 2010*. The system achieved an average accuracy of 80% for all languages to Hindi and English. However, an analysis of the erroneous transliterations revealed that either the provided test transliteration was wrong or the transliteration contained some not so frequently used characters from the language. From our observation, we found that the system performs much more accurately for regular search queries.

D. Overall System Performance for Telugu CLIA

Overall system was evaluated using the P@5 and P@10 scores for 5 queries .Table 3 below shows the scores for each of the queries.

TABLE II. P@5 AND P@10 FOR 5 QUERIES FROM TELUGU-HINDI

Query No.	P@5	P@10
1	0.6	0.5
2	0.6	0.6
3	0.4	0.4
4	0.4	0.3
5	0.8	0.8

TABLE III. P@5 AND P@10 FOR 5 QUERIES FROM TELUGU-ENGLISH

Query No.	P@5	P@10
1	0.4	0.6
2	0.4	0.3
3	0.4	0.4
4	0.8	0.5
5	0.8	0.9

VII. Conclusion and Future work

This paper gave a basic introduction to what a Cross Lingual Information Access system and why such a system is necessary for the Web. This work also covers the various issues that are to be dealt with for Indian Languages in particular in the context of the CLIA project.

The paper gave an overall architecture of the CLIA system with detailed descriptions and experimental results for the Language Identification, Domain Identification and Query Processing tasks that form the core of the CLIA system.

Continuous effort is going into the project to improve the performance of the system at various levels. As time progresses, we plan to categorize the web into more and more domains for effective and efficient information extraction. Currently the system supports cross-retrieval from Hindi and English given a query in any of the supported Indian Languages. This will soon change as we design a query processing system that can formulate queries from any supported language to any other supported language.

Acknowledgment

We acknowledge the support we have received from Ministry of Communication and Information Technology (MCIT), Government of India by funding the project. We also acknowledge the support given by rest of the Cross Language Information Access (CLIA) consortium members by providing us the necessary language resources.

References

[1] Oard, D., & Dorr, B. (1996). *A Survey of Multilingual Text Retrieval*. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies.

[2] Larkey, L., AbdulJaleel, N., & Connell, M. (2003).*What's in a Name? : Proper Names in Arabic Cross Language Information Retrieval.* CIIR Technical Report, IR-278, Univ. of Amherst.

[3] Saravanan, K., Udupa, R., & Kumaran, A. (2010). *Cross lingual Information Retrieval System Enhanced with Transliteration Generation and Mining*. In Proceedings of Forum for Information Retrieval Evaluation (FIRE-2010) Workshop, Kolkata, India.

[4] Ea-Ee Jan, Shih-Hsiang Lin, & Berlin Chen. (2010). *Transliteration Retrieval Model for Cross Lingual Information Retrieval.* Lecture Notes in Computer Science, 2010, Volume 6458, Information Retrieval Technology, pp.183-192.

[5] Sanderson, M. (1994).*Word sense disambiguation and information retrieval*. In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Dublin, Ireland, p.142-151.

[6] Chakrabarti, S., Dom, B., & van den Berg, M. (1999). *Focused Crawling: A New Approach for Topic-Specific Resource Discovery*. In Proc. 8th World Wide Web Conf., Elsevier Science, Amsterdam, pp. 545-562.

[7] Pingali, P., & Varma, V. (2006).*Hindi and Telugu to English Cross Language Information Retrieval* .Working Notes of Cross Language Evaluation Forum Workshop, Spain.

[8] Pingali P., Kula K., T., & Varma, V. (2007). *Hindi, Telugu, Oromo, English CLIR Evaluation.* Evaluation of Multilingual and Multi-modal Information Retrieval. Vol. 4730, 2007, ISBN 978-3-540-74998-1, Springer-Verlag.

[9] Pingali, P., & Varma, V. (2007). *Multilingual Indexing Support for CLIR using Language Modeling.* In Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.

[10] Pingali, P., Jagarlamudi, J., & Varma, V. (2006). *A Dictionary Based Approach with Query Expansion to Cross Language Query Based Multi-Document Summarization: Experiments in Telugu English.* National Workshop on Artificial Intelligence, Mumbai, India.

[11] Wallach H., M. (2004). *Conditional random fields: An introduction*. Technical Report MS-CIS-04-21, University of Pennsylvania, Department of Computer and Information Science, University of Pennsylvania.

[12] Shishtla, P., M., Pingali, P., & Varma, V. (2008). *A Character n-gram Based Approach for Improved Recall in Indian Language NER.* In Proceedings of IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages, Hyderabad, India, pp 67-74.

[13] Shishtla, P., M., Pingali, P., & Varma, V. (2008). *Experiments in Telugu NER: A Conditional Random Field Approach*. In Proceedings of IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages, Hyderabad, India, pp 105-110.

[14] Ballesteros, L., & Croft W., B. (1997). *Phrasal translation and query expansion techniques for cross-language information retrieval.* In Proceedings of the 20th Annual International ACM Conference on Research and Development in Information Retrieval, Philadelphia, PA, July 27–31, pp.84–91.

[15] Robertson, S., & Zaragoza, H. (2009). *Probabilistic Relevance Framework: BM25 and Beyond.* Foundations and Trends in Information Retrieval. pp.333-389.

[16] Saracevic, T. (1975). *Relevance: A review of and a framework for thinking on the notion in information science*. Journal of the American Society for Information Science and Technology, pp. 321-343.