

Experiments in Cross-lingual IR among Indian Languages

Ranbeer Makin, Nikita Pandey, Prasad Pingali and Vasudeva Varma

International Institute of Information Technology, Hyderabad, India

Abstract. Commonly used vocabulary in Indian language documents found on the web contain a number of words that have Sanskrit, Persian or English origin. However, such words may be written in different scripts with slight variations in spelling and morphology. In this paper we explore fuzzy string matching techniques to exploit this situation of relatively large number of cognates among Indian languages, which are higher when compared to an Indian language and a non-Indian language. We present an approach to identify cognates and make use of them for improving dictionary based CLIR when the query and documents both belong to two different Indian languages. We conduct experiments using a Hindi document collection and a set of Telugu queries and report the improvement due to cognate recognition and translation.

Key words: Telugu-Hindi CLIR, Indian Languages, Cognate Identification

1 Introduction

India is a multi-language, multi-script country with 22 official languages and 11 written script forms. About a billion people use these languages as their first language. A huge amount of regional news and cultural information is usually found on the web in these languages and is inaccessible to people of other regions within the country. Information access technologies such as Cross-Language Information Retrieval (CLIR) across various Indian languages remain largely unexplored. All previous CLIR research involving Indian languages were conducted in combination with English. For example, ACM TALIP¹ conducted a surprise language exercise in 2003, which focused on CLIR systems to retrieve Hindi documents for the given English queries. Similarly, ad-hoc CLIR evaluation tasks were conducted at CLEF² in 2006 to evaluate systems' [1] performance to retrieve English documents for a given set of Hindi and Telugu queries. Most of the Indian language texts in the print and online media have a number of words that have originated from Sanskrit, Persian and English. While in many cases one might argue that such occurrences do not belong to an Indian language, the frequency of such usage indicates a wide acceptance of these foreign language

¹ ACM Transactions on Asian Language Information Processing.

² Cross Language Evaluation Forum. <http://www.clef-campaign.org>

words as Indian language words. In many cases these words also are morphologically altered as per the Indian language morphological rules to generate new variant words. We treat all such words which have a common origin as *cognates* and study how we can use fuzzy string matching techniques to the problem of CLIR. In this paper we particularly attempt to exploit the similarity among various Indian language words, which may share relatively more number of cognates when compared to an Indian language and another non-Indian language. We focus on identifying cognates between pairs of Indian languages and use them to make the retrieval more effective. An example of a cognate pair for the word ‘school’ in English, across Indian languages is ‘विध्यालय’ (pronounced as ‘vidhyaalaya’) in Hindi and ‘విద్యాలయము’ (pronounced as ‘vidyaalayamu’) in Telugu, both of which are derived from Sanskrit. Cognate identification has been found to be useful in aligning sentences [2], aligning words [3], and in translation lexicons induction [4,5]. In CLIR, Pirkola et al. [6] extracted similar terms between English and Spanish from a bilingual dictionary to assist in automatic rule generation for translation and many studies similar to these exist in closely related languages. However, no such studies exist to study the effect of cognates in CLIR when the documents are to be retrieved from one Indian language for a given query in a different Indian language. In this paper we conduct some experiments in this direction and explore some fuzzy string matching techniques and their performance in the context of Indian language CLIR.

The paper is organized as follows. Section 2 brings forth the major issues in CLIR and discusses the motivation for a cognate-based approach for Indian language CLIR. Section 3 gives a detailed description of the Indian language to Indian language CLIR system architecture. In Section 4, we discuss the evaluation framework for our system, the setup and results of our experiments. Finally, in Section 5, we discuss the future work and conclude. But no studies exist to study similar approaches for Indian language to Indian language CLIR.

2 Issues in CLIR

The problem of CLIR is defined as retrieval of relevant documents of language L_2 for a given query of language L_1 . Some of the key technical issues [7] for CLIR can be thought of as

1. How can a query term in L_1 be expressed in L_2 ?
2. What mechanisms determine which of the possible translations of text from L_1 to L_2 should be retained?
3. In cases where more than one translation are retained, how can different translation alternatives be weighed?

Among these issues, the first issue forms the most fundamental issue that needs to be addressed before others can be attempted. The first issue mentioned above primarily deals with obtaining a query translation into the document’s language. The three basic approaches to address this issue include: machine translation (MT), parallel or comparable corpus and machine-readable bilingual

dictionary. However, each of them has one or the other drawbacks. In the case of machine translation, the search queries are unable to impart structural information to MT systems, needed by them to perform disambiguation in translation. Hence, the quality of translation by MT systems is poor [8, 9]. The parallel corpus approach is less effective because the topic coverage of such a corpus is limited [10].

Bilingual dictionaries generally contain more verbose definitions with examples which are not very suitable for retrieval. An IR system needs only direct translation of each search term [8]. In general, proper names and technical terms are absent in these dictionaries used by CLIR systems. The corresponding expressions in the target language are sometimes identical or can be found by transliteration. However, they often differ with slight spelling or morphological variations in the target language [6]. Also, a bilingual dictionary from a source language to a target language generally has a greater coverage of source language words compared to those of target language to source language. Thus, using only a bilingual dictionary approach can miss out on some of the words of the target language that might have been present in the target documents. These issues of CLIR also apply in Indian language to Indian language (IL-IL) information retrieval scenario as well. As Indian languages exhibit significant similarity in vocabulary, we have incorporated cognate identification technique in addition to using a bilingual dictionary for word-word translation to deal with the first issue mentioned above. However, we also partially address the second issue of determining the number of cognate translations to be retained from all possible translations resulting from our algorithm.

3 Indian Language CLIR System Architecture

In this paper, we report an Indian language - Indian language information retrieval system which takes a query in one Indian language (IL1) and retrieves documents of another Indian language (IL2). The high-level architecture of this system is depicted in Figure 1.

The user issues a query in IL1 which is tokenized into keywords. These query keywords are then looked up in IL1-IL2 bilingual dictionary to get the corresponding IL2 keywords.

Cognate Identification. The IL1 query keywords are also searched for their corresponding cognates in IL2. For this, we first extract words from an IL2 corpus to have a reasonably good vocabulary of IL2. This corpus has been collected by crawling Hindi web sources over a period of time. For each query keyword in IL1, the IL2 vocabulary list is searched to identify its cognates. We hypothesize that the likelihood of the two words across a pair of Indian languages to be cognates is highly correlated with their orthographic similarity. Hence we use the string similarity metrics for cognate identification. In this work, we make use of the *Jaro-Winkler* similarity [11] [12], which adjusts the weights of pairs s , t that share a common prefix to give them more favorable score; the *Levenstein*

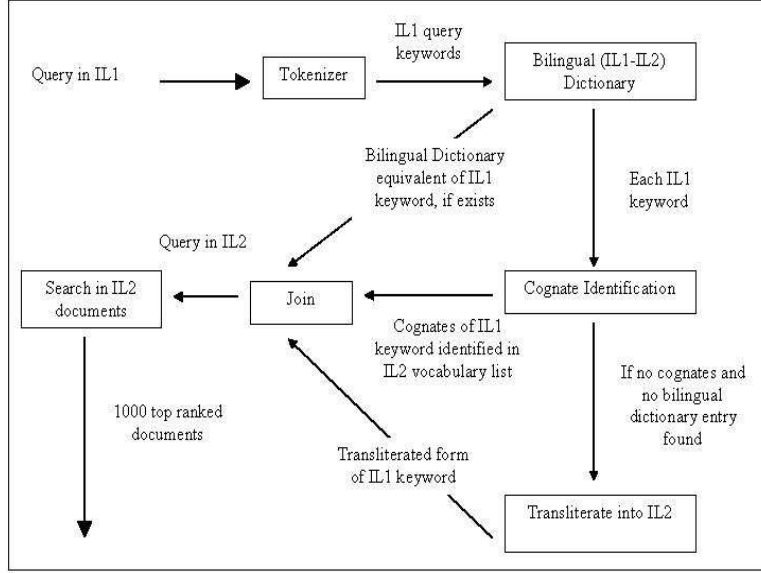


Fig. 1. High-level architectural view of Indian language - Indian language CLIR system.

distance, which is a string similarity measure, defined as the minimum cost needed to convert a string s into another string t ; and the *Longest Common Subsequence Ratio*, or LCSR [2] which takes the ratio of the longest common subsequence of pairs s, t to the length of the longest string amongst the two. Jaro-Winkler's similarity score is computed as follows:

$$JaroWinkler(s, t) = Jaro(s, t) + \left(\frac{P}{10} * (1.0 - Jaro(s, t)) \right)$$

$$Jaro(s, t) = \frac{1}{3} \left(\frac{|s'|}{|s|} + \frac{|t'|}{|t|} + \frac{|s'| - T_{s',t'}}{2|s'|} \right)$$

where, P: length of common prefix; s, t : input strings; s' : characters in s that are common with t ; t' : characters in t that are common with s ; $T_{s',t'}$: number of transpositions of characters in s' relative to t .

And LCSR is defined as:

$$LCSR(s, t) = \frac{|LCS(s, t)|}{\max(|s|, |t|)}$$

where, s, t are two strings; $LCS(s, t)$ is the longest common subsequence in s and t .

Since the scripts of IL1 and IL2 may differ, our cognate identification technique performs a phonetically motivated comparison of IL1 and IL2 words using the above mentioned orthographic similarity functions. This phonetic based approach allows matching to be carried out across any pair of two scripts.

The keywords, for which no bilingual dictionary equivalents and no cognates are identified, are transliterated into IL2 using a pre-determined set of mapping rules between the two scripts. The combined query resulting from all these three steps, viz. bilingual dictionary look-up, cognate identification, and transliteration, is then used to retrieve the IL2 documents using the full-featured text search engine, *Lucene*³. The result set of documents obtained is ranked according to Lucene's scoring criterion⁴ from which only 1000 highest-ranked documents are collected.

4 Experiments

The experiments were carried out on the two Indian languages, Hindi and Telugu. Though our CLIR system can work across any pair of Indian languages, the choice of the above two languages was made to ease the relevance judgment and manual translation tasks of the authors. The document collection for our experiments comprised of electronic news articles (in Hindi) published during 2003 and 2006 by the websites *BBC Hindi*⁵ and *Navbharat Times*⁶. These documents covered various domains including politics, sports, science, entertainment, etc. The test set consisted of 50 Telugu queries.

4.1 Evaluation Framework

We came up with an evaluation framework to assess the performance of our Indian Language CLIR system. As the first step towards building this framework, the test queries were manually translated to Hindi. Manual translation has the drawback of involving human judgment; nevertheless, being an inexpensive and a much effective strategy, it was preferred to machine translation. Relevance judgment was then manually performed for each of these Hindi queries to facilitate direct comparisons between the cross-lingual performance of our system with its monolingual performance.

4.2 Setup

In our work, we experimented with the Jaro-Winkler, Levenstein distance and LCSR similarity measures individually to identify cognates. The binary classification of cognates was done with an empirically chosen threshold⁷. The list of

³ Text Search Engine Lucene - <http://lucene.apache.org/>

⁴ <http://lucene.apache.org/java/docs/api/org/apache/lucene/search/Similarity.html>

⁵ <http://www.bbc.co.uk/hindi/>

⁶ <http://navbharattimes.indiatimes.com/>

⁷ Thresholds chosen were 0.90 for Jaro-Winkler, and 0.85 for Levenstein and LCSR.

potential Telugu-Hindi cognate pairs thus obtained was sorted in the descending order of the scores assigned by the similarity functions. We believe that the true cognates will occur more frequently towards the top of the sorted list and decrease in frequency as we descend this list. Based on this belief, we introduced the notion of *window size*, which defines the number of cognates to be taken for every Telugu keyword. The experiments were conducted with window size varying from 1 to 10, where the maximum limit was empirically chosen.

4.3 Results

Experiments were performed with seven models, where the first model uses bilingual dictionary⁸ alone. The next three models (Jaro-Winkler, LCSR, and Levenstein) are based on orthographic similarity, and perform cross-language retrieval exclusively on the basis of cognates identified. The last three models combine the bilingual dictionary approach with each of the cognate identification techniques. We evaluated our experimental results on 11-point interpolated recall - precision averages [13], mean average precision, geometric average precision, and recall using standard *trec-eval*.

In this section, we compare the seven different models and analyze the performance of our CLIR system with each of these models. We then discuss the effect of varying window size on the performance of a model.

Comparisons. Table 1 compares recall, mean average precision (MAP), and geometric average precision (GAP) of seven different models, for window size 3 (the performance of our system was comparatively better on this window size), on the test set of 50 Telugu queries. Model-1 is the bilingual dictionary approach and is chosen as the baseline method. Model-2 is based on only the cognate identification approach using Jaro-Winkler similarity. Similarly, Model-3 corresponds to LCSR, and Model-4 to Levenstein distance. Model-5 to Model-7 combine bilingual dictionary approach with Model-2 to Model-4 respectively.

	Model-1	Model-2	Model-3	Model-4	Model-5	Model-6	Model-7
Recall	0.6059	0.5381	0.4865	0.4479	0.6875	0.6628	0.6418
Mean Avg Precision	0.1647	0.2498	0.1976	0.1692	0.2771	0.2449	0.2074
Geometric Avg Precision	0.0048	0.0126	0.0042	0.0023	0.0263	0.0186	0.0113

Table 1. Comparison of recall, mean average precision, and geometric average precision for all the models on window size 3.

Surprisingly, impressive results are achieved with the cognate techniques alone. Cross-lingual retrieval based only on the cognates identified using Jaro-Winkler

⁸ Telugu-Hindi bilingual dictionary <http://ltrc.iitit.net/onlineServices/Dictionaries/Tel-Hin.DictDwnld.html>

similarity shows an increase of 51.67% in mean average precision and 162.5% in geometric average precision, with only a slight decrease of 11.2% in recall.

Table 1 also strongly suggests that combining the bilingual dictionary approach with the cognate identification techniques in Indian language - Indian language scenario yields more effective results than using these approaches individually. This is not unexpected as the drawbacks of taking only the dictionary approach, as mentioned in Section 2, are solved to a good extent by using cognates. Similarly, only cognate techniques do not perform as well as the combined approaches since there is a possibility that cognate pairs can have different meanings. Also due to partial overlap in the vocabulary of Indian languages, cognates may not necessarily exist for every word. These drawbacks are compensated by the use of bilingual dictionary.

Even among the combined approaches, dictionary with Jaro-Winkler similarity computation shows better performance than the other two. For window size 3, we observe that this model leads to a significant increase of 68.25% in mean average precision, 447.92% in geometric average precision and 13.45% in recall.

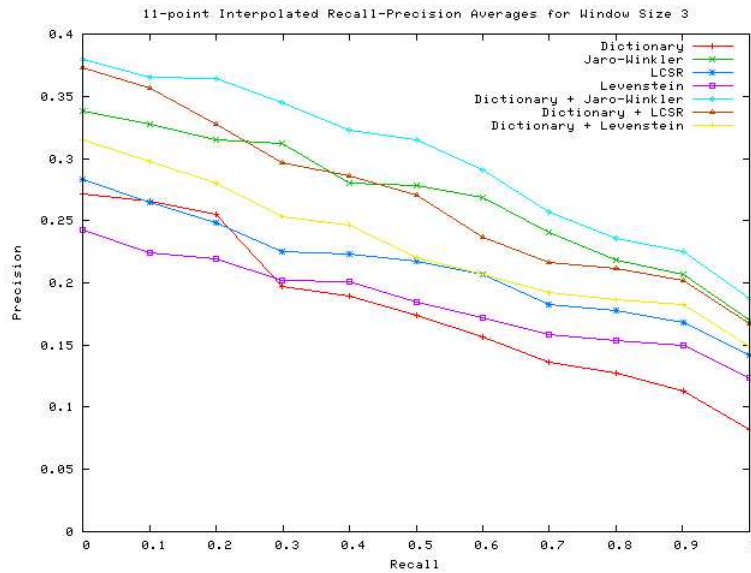


Fig. 2. 11-point interpolated recall precision curves for all the models on window size 3. The x-axis represents various recall levels and y-axis represents interpolated precision at these levels.

Figure 2 gives a more detailed comparison of the effectiveness of the models on test queries for window size 3, in the form of 11-point interpolated recall-precision curves. These curves confirm to our findings above. The variations in

the results obtained on varying the similarity measures are highly correlated to how well the cognates are identified by these measures.

Window Size Variation. Table 2 shows the effect of variations in window size on the combined approach of dictionary and LCSR. We notice that significant variations in recall, mean average precision, and geometric average precision occur when the window size is varied from 1 to 3. The variations in these measures decrease as the window size is further varied from 4 to 6. On any further increase in the window size, we observe that the variations become more or less constant. This suggests that the maximum number of true cognates get identified within window size 3, which confirms to our belief that true cognates occur near the top of the sorted cognate pairs list. Similar behavior is observed for other models as well.

Window Size	1	2	3	4	5	6	7	8	9	10
Recall	0.6860	0.6512	0.6628	0.6574	0.6767	0.6744	0.6775	0.6775	0.6775	0.6775
MAP	0.2313	0.2439	0.2449	0.2435	0.2336	0.2354	0.2387	0.2405	0.2404	0.2414
GAP	0.0168	0.0167	0.0186	0.0173	0.0188	0.0189	0.0213	0.0212	0.0211	0.021

Table 2. Effect of varying window size from 1 to 10 on recall, mean average precision (MAP), and geometric average precision (GAP) using the bilingual dictionary approach with LCSR.

5 Conclusion and Future Work

We came up with an Indian language - Indian language information retrieval system, which exploits the significant overlap in vocabulary across the Indian languages. We identified cognates using some of the well-known similarity measures, and incorporated this technique with the traditional bilingual dictionary approach. The effectiveness of our retrieval system was compared on various models. The results show that using cognates with the existing dictionary approach leads to a significant increase in the performance of our system. Experiments have also led to the surprise finding that our Indian Language CLIR system based only on the cognates approach performs better than the dictionary approach alone. This shows a good promise for cross-lingual retrieval across those pairs of related languages for which bilingual dictionaries do not exist.

In the future, we would like to measure the degree of similarity among other Indian languages with our CLIR system. We would also like to extend our system to perform cross-lingual retrieval across those pairs of Indian languages which have a little overlap between their vocabularies, but are significantly related to some third Indian language.

References

1. Pingali, P., Varma, V.: Hindi and Telugu to English Cross Language Information Retrieval at CLEF 2006. In: Working Notes of Cross Language Evaluation Forum 2006. (2006)
2. Melamed, I.D.: Bitext maps and alignment via pattern recognition. *Computational Linguistics* **25**(1) (1999) 107–130
3. Tiedmann, J.: Combining clues for word alignment. In: Proceedings of the 10th Conference of the European Chapter of the ACL (EACL'03). (2003)
4. Koehn, P., Knight, K.: Knowledge sources for word-level translation models. In: Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing. (2001) 27–35
5. Mann, G.S., Yarowsky, D.: Multipath translation lexicon induction via bridge languages. In: Proceedings of NAACL 2001. (2001) 151–158
6. Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K., Jarvelin, K.: Fuzzy translation of cross-lingual spelling variants. In: Proceedings of SIGIR'03. (2003) 345–352
7. Grefenstette, G., Grefenstette, G.: *Cross-Language Information Retrieval*. Kluwer Academic Publishers, Norwell, MA, USA (1998)
8. Hull, D., Grefenstette, G.: Querying across languages: A dictionary-based approach to multilingual information retrieval. In: Proceedings of the 19th Annual international ACM SIGIR 1996, Zurich, Switzerland (1996) 49–57
9. Radwan, K., Fluhr, C.: Textual database lexicon used as a filter to resolve semantic ambiguity application on multilingual information retrieval. In: The 4th Symp. On Document Analysis and Information Retrieval. (1995) 121–136
10. Adriani, M., Croft, W.: The effectiveness of a dictionary-based technique for indonesian-english cross-language text retrieval. CLIR Technical Report IR-170 (1997)
11. Jaro, M.: Probabilistic linkage of large public health data files. *Statistics in Medicine* **14** (1995) 491–498
12. Winkler, W.: The state record linkage and current research problems. Technical report, statistics of Income Division, Internal Revenue Service Publication (1999)
13. Manning, C.D., Schütze, H.: *Foundations of Statistical Natural Language Processing*. The MIT Press (2001)