Approximate String Matching Techniques for Effective CLIR Among

by

Ranbeer Makin, Nikita Pandey, Prasad Pingali, Vasudeva Varma

in

International Workshop on Fuzzy Logic and Applications, Ruta di Camogli, Genova, Italy - July 7-10, 2007, F. Masulli, S. Mitra, and G. Pasi (Eds.): WILF 2007, LNAI 4578, pp. 430-437, 2007. Springer-Verlag Berlin Heidelberg 2007

Report No: IIIT/TR/2008/40

Centre for Search and Information Extraction Lab
International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2008

Approximate String Matching Techniques for Effective CLIR among Indian Languages

Ranbeer Makin, Nikita Pandey, Prasad Pingali and Vasudeva Varma

International Institute of Information Technology,
Hyderabad, India
{ranbeer,nikita}@students.iiit.ac.in, {pvvpr,vv}@iiit.ac.in

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 approximate 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 TALIP[1] 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 CLEF[2] in 2006 to evaluate systems' performance to retrieve English documents for a given set of Hindi and Telugu queries [1]. 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

[1] ACM Transactions on Asian Language Information Processing.
[2] Cross Language Evaluation Forum. *http://www.clef-campaign.org*

of such usage indicates a wide acceptance of these foreign language words as Indian language words. In many cases these words are also 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 approximate string matching techniques to the problem of CLIR. An example of a cognate pair for the word 'school' in English, across Indian languages is '**विद्यालय**' (pronounced as 'vidyaalaya') in Hindi and 'విద్యాలయము ' (pronounced as 'vidyaalayamu') in Telugu, both of which are derived from Sanskrit. 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.

Some of the traditional approaches to perform query translation for CLIR include machine translation (MT), parallel or comparable corpus and machine-readable bilingual dictionary. MT and parallel corpus based approaches do not work well, in general, for CLIR [2,3] [4]. 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 [2]. In general, proper names and technical terms are absent in these dictionaries used by CLIR systems. Also, a bilingual dictionary has a greater coverage of source language words compared to that of target 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 documents. These issues of CLIR also apply in Indian language to Indian language (IL-IL) information retrieval scenario. As Indian languages exhibit significant similarity in vocabulary, we incorporate cognate identification technique in addition to using a bilingual dictionary.

Cognate identification has been found to be useful in aligning sentences [5], aligning words [6], and in translation lexicons induction [7, 8]. In CLIR, Pirkola et al. [9] 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 approximate string matching techniques and their performance in the context of Indian language CLIR.

The paper is organized as follows. Section 2 gives a detailed description of the Indian language to Indian language CLIR system architecture. In Section 3 we describe the evaluation framework of our system and experimental setup, and in Section 4 we present the results of our experiments. Finally, in Section 5, we discuss the future work and conclude.

2 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

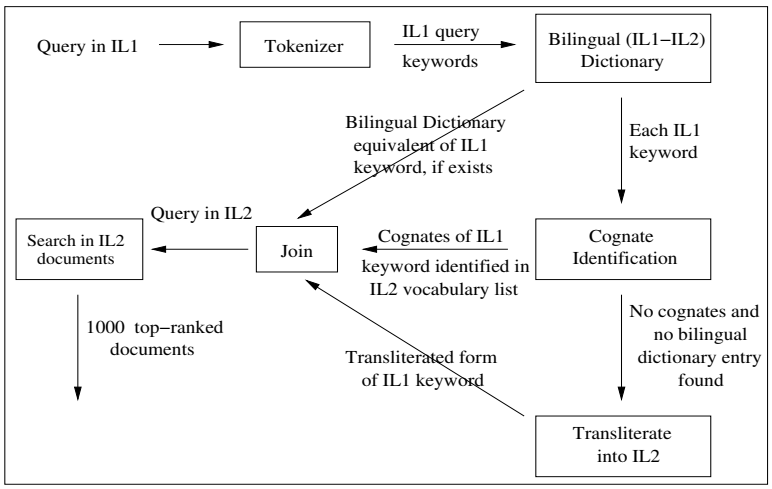

Fig. 1. High-level architectural view of Indian language - Indian language CLIR system.

into keywords. These query keywords are then looked up in IL1-IL2 bilingual dictionary to get the corresponding IL2 keywords.
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. 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 [10] [11], which adjusts the weights of pairs s, t that share a common prefix to give them more favorable score; the *Levenstein* 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 [5] 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:

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

$$Jaro(s,t) = \frac{1}{3}\left(\frac{|s\prime|}{|s|} + \frac{|t\prime|}{|t|} + \frac{|s\prime| - T_{s\prime,t\prime}}{2|s\prime|}\right)$$

where P: length of common prefix; s, t: input strings; $s\prime$: characters in s that are common with t; $t\prime$: characters in t that are common with s; $T_{s\prime,t\prime}$: number of transpositions of characters in $s\prime$ relative to $t\prime$.
And LCSR is defind as:

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

where LCS(s,t) is the longest common subsequence in strings 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 criterion from which only 1000 highest-ranked documents are collected.

3 Experiments

The experiments were carried out on the two Indian languages, Hindi and Telugu. The choice of the above two languages was made because of the availability of the resources, and to ease the relevance judgment and manual translation tasks. As Hindi and Telugu are quite different in nature and the cognate identification technique is independent of the languages used, our system can work across any pair of Indian languages. The document collection for our experiments comprised of around 50,000 electronic news articles (in Hindi) published during 2003 and 2006 by the *BBC Hindi* and *Navbharat Times* websites[4]. These documents covered various domains including politics, sports, science and entertainment. The test set consisted of 50 Telugu queries framed by the native Telugu speakers, based on the guidelines that the queries should be related to the events occurred during 2003 and 2006, and should belong to the above mentioned domains.

Evaluation Framework. We used Cranfield evaluation methodology to assess the performance of our Indian Language CLIR system. Relevance judgment was manually performed by the native Hindi speakers, for which the Telugu test queries had to be translated into Hindi. These Hindi speakers were different from the people who came up with the test set.

[3] Text Search Engine Lucene - *http://lucene.apache.org/*
[4] *http://www.bbc.co.uk/hindi/* and *http://navbharattimes.indiatimes.com/*

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 threshold[5]. The list of 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.
Experiments were performed with six models, where the first 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. To compare the performance of our system, two baseline methods were chosen, the upper baseline being the monolingual performance of our system and the lower one being the bilingual dictionary[6] method.

4 Results

We evaluated our experimental results on 11-point interpolated recall - precision averages [12], mean average precision (MAP), geometric average precision (GAP), and recall using standard *trec-eval*. Baseline-1 is the monolingual performance of the system and Baseline-2 is the bilingual dictionary approach. Model-1 is based on only the cognate identification approach using Jaro-Winkler similarity. Similarly, Model-2 corresponds to LCSR, and Model-3 to Levenstein distance. Model-4 to Model-6 combine bilingual dictionary approach with Model-1 to Model-3 respectively. In this section, we compare the six 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. The Hindi monolingual run retrieved relevant documents for all the 50 queries. However, relevant documents were retrieved by only 72% (36) of the queries in the cross-lingual run using Baseline-2. Table 1 compares recall, MAP, and GAP of six different models for window size 3 (the performance of our system was comparatively better on this window size) with the baselines on the test set of 50 Telugu queries.
Surprisingly, impressive results are achieved with the cognate techniques alone. Cross-lingual retrieval based only on the cognates identified using Jaro-Winkler similarity shows an increase of 51.67% in MAP and 162.5% in GAP on comparison with Baseline-2, with only a slight decrease of 11.2% in recall.

[5] Thresholds chosen were 0.90 for Jaro-Winkler, and 0.85 for Levenstein and LCSR.
[6] Telugu-Hindi bilingual dictionary *http://ltrc.iiit.net/onlineServices/Dictionaries/Tel-Hin_DictDwnld.html*

	Baseline-1	Baseline-2	Model-1	Model-2	Model-3	Model-4	Model-5	Model-6
Recall	0.9907	0.6059	0.5381	0.4865	0.4479	0.6875	0.6628	0.6418
MAP	0.5611	0.1647	0.2498	0.1976	0.1692	0.2771	0.2449	0.2074
GAP	0.4133	0.0048	0.0126	0.0042	0.0023	0.0263	0.0186	0.0113

Table 1. Comparison of recall, MAP, and GAP for all the models on window size 3 and the baselines.

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 1, 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 by and large compensated by the use of bilingual dictionary.

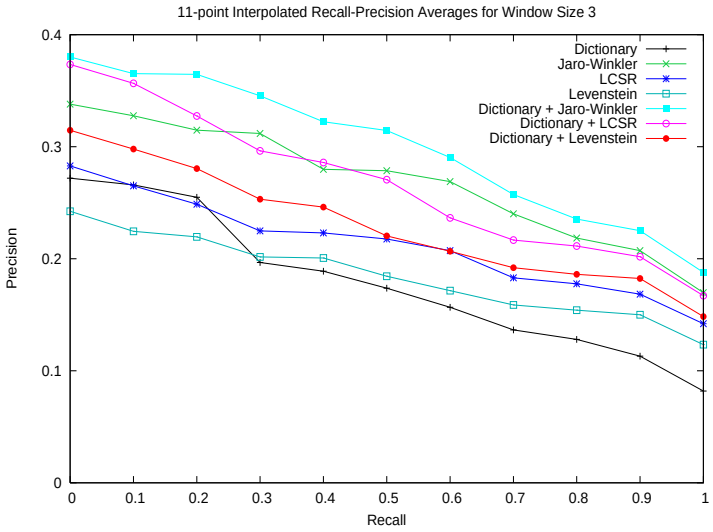

Fig. 2. 11-point interpolated recall precision curves for all the models on window size 3. X-axis represents various recall levels and Y-axis represents interpolated precision at these levels.

Even among the combined approaches, dictionary with Jaro-Winkler similarity computation shows better performance than the other two combined approaches. Using this Model-4, on an average, only 84% (42) of the queries retrieved relevant

documents. 78.57% (33) of these 42 queries performed better with this model in terms of recall and average precisions than with Baseline-2. We believe that good cognates couldn't be found for the keywords in the remaining 21.43% (9) of the queries, and hence the performance deteriorated. Out of the 14 queries for which Baseline-2 could not retrieve relevant documents, on an average Model-4 succeeded in retrieving for 50% of them. For window size 3, we observe that this model leads to a significant increase of 68.25% in MAP, 447.92% in GAP and 13.45% in recall on comparison with Baseline-2. However, its performance is still lower by 30.60% in recall, 50.61% in MAP, and 93.64% in GAP when compared with Baseline-1. This is very much expected as the monolingual retrieval performance is generally acknowledged as the practical limit.
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 apporach of dictionary and LCSR. We notice that significant

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, MAP, and GAP using the bilingual dictionary approach with LCSR.

variations in recall, MAP, and GAP 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.

5 Conclusion and Future Work

We came up with an Indian language - Indian language IR 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, on an average, 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. 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
3. 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
4. Adriani, M., Croft, W.: The effectiveness of a dictionary-based technique for indonesion-english cross-language text retrieval. CLIR Technical Report IR-170 (1997)
5. Melamed, I.D.: Bitext maps and alignment via pattern recognition. Computational Linguistics **25(1)** (1999) 107–130
6. Tiedmann, J.: Combining clues for word alignment. In: Proceedings of the 10th Conference of the European Chapter of the ACL (EACL'03). (2003)
7. 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
8. Mann, G.S., Yarowsky, D.: Multipath translation lexicon induction via bridge languages. In: Proceedings of NAACL 2001. (2001) 151–158
9. 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
10. Jaro, M.: Probabilistic linkage of large public health data files. Statistics in Medicine 14 (1995) 491–498
11. Winkler, W.: The state record linkage and current research problems. Technical report, statistics of Income Division, Internal Revenue Service Publication (1999)
12. Manning, C.D., Schutze, H.: Foundations of Statistical Natural Language Processing. The MIT Press (2001)