Experiments in CLIR Using Fuzzy String Search Based on Surface Similarity

Sethuramalingam S, Anil Kumar Singh, Pradeep Dasigi and Vasudeva Varma
Language Technologies Research Centre
International Institute of Information Technology
Hyderabad, India
{sethu|anil}@research.iiit.ac.in, pradeep.dasigi@gmail.com, vv@iiit.ac.in

ABSTRACT

Cross Language Information Retrieval (CLIR) between languages of the same origin is an interesting topic of research. The similarity of the writing systems used for these languages can be used effectively to not only improve CLIR, but to overcome the problems of textual variations, textual errors, and even the lack of linguistic resources like stemmers to an extent. We have conducted CLIR experiments between three languages which use writing systems (scripts) of Brahmi-origin, namely Hindi, Bengali and Marathi. We found significant improvements for all the six language pairs using a method for fuzzy text search based on Surface Similarity. In this paper we report these results and compare them with a baseline CLIR system and a CLIR system that uses Scaled Edit Distance (SED) for fuzzy string matching.

Categories and Subject Descriptors

H.3.2 [**Information Storage and Retrieval**]: Query formulation; H.3.1 [**Content Analysis and Indexing**]: Linguistic processing

General Terms

Algorithms, Experimentation, Measurement, Performance

Keywords

Surface Similarity, Fuzzy text search, Spelling variations

1. INTRODUCTION

CLIR systems need to be robust enough to tackle textual variations or errors both at the query end and at the document end. There are various reasons for textual variations like spelling variations, dialectal variations, morphological variations etc. Some of the common causes of errors are typographical errors and errors in font conversion across multiple proprietary encoding formats. IR systems have to take care of such variations [6, 2] to be able to perform reasonably well.

One of the techniques to achieve this is to use approximate or fuzzy string matching [5]. If such a technique benefits from the knowledge about the characteristics of the writings systems and their similarities, then it can further help in cross-lingual IR. In this paper, we report the results of using one such technique of implementing fuzzy text search for CLIR to solve the above mentioned problems. We have used it for three popular languages (Hindi, Bengali and Marathi) which use Brahmi origin scripts. Our method of fuzzy text search could be used in any type of CLIR system irrespective of their underlying retrieval models.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

2. SURFACE SIMILARITY ESTIMATION

Surface Similarity is a general term for orthographic and phonetic similarity (i.e., similarity of 'form') between any two words of a language. Singh [5] defined it as

$$S = \psi(w_1,\ w_2,\ A,\ O,\ O_n,\ P,\ P_n,\ D) \qquad (1)$$

where S is the surface similarity of two words w_1 and w_2 and ψ is the function that estimates this similarity in terms of the Model of Alphabet A, the sets of orthographic features and their numeric values O and O_n respectively, the sets of phonetic features and their numeric values P and P_n respectively, and D, the Stepped Distance Function (SDF) for calculating the similarity between any two letters. This formulation of Surface Similarity leaves open the way in which the function ψ is implemented and also the way in which the parameters are modeled. For our experiments, we use the modified Dynamic Time Warping (DTW) algorithm as an implementation of this function and use the Computational Phonetic Model of Scripts or CPMS [4] for modeling the parameters. The DTW algorithm aligns the two strings based on the CPMS for Brahmi origin scripts.

3. CLIR BASED ON FUZZY TEXT SEARCH

In the general case, CLIR would require proper query translation. However, in the case of related languages, especially when the required resources are not available or are insufficient, CLIR can still be achieved to a reasonable degree through fuzzy or approximate text search based on Surface Similarity. This will work for related languages because these languages are likely to have a large number of words of common origin and also because they use the same or similar writing systems. Retrieval is performed in such a case by converting the query to the writing system of the target language (a relatively simple task for very similar writing systems) and then applying fuzzy string matching while matching words (to take care of variations etc.). For domain specific retrieval, this simple method can be even more effective, as was the case with our experiments.

4. EVALUATION

For evaluation, we used the CLIR data released at the FIRE[1] workshop, 2008. The corpora consisted of comparable news articles in Hindi, Bengali, and Marathi collected during 2004 to 2007. We used the 50 test queries provided at the FIRE workshop in each of these languages for our evaluation.

We conducted experiments on six different pairs of these three languages. The Corpora were indexed using the open-source search engine, Lucene[2]. Lucene's OKAPI BM25 was used as the similarity metric for scoring and ranking the documents. The Run IDs and their descriptions are shown in Table 1. The fuzzy text search implementation in our system takes the input query word from the source language and returns the closest fuzzy string match from the target language. For reasons of convenience, we used Devanagari as the common script (to which all data was converted) for the experiments.

A simple, dictionary-based approach was used as the lower baseline for comparison. We also compared the performance of the fuzzy text search with Scaled Edit Distance or SED [1] based search for each of these runs. The reason for preferring SED over the normal edit distance is that it reduces the disparities between long and short words. The comparison of results is shown in Table 2. To estimate the importance of Surface Similarity, we explicitly avoided using any stemmers or morphological analyzers.

Table 1: Runs Descriptions

Run ID	Description
BH	Bengali queries and Hindi documents
BM	Bengali queries and Marathi documents
HB	Hindi queries and Bengali documents
HM	Hindi queries and Marathi documents
MB	Marathi queries and Bengali documents
MH	Marathi queries and Hindi documents

To evaluate the significance of our retrieval results, we conducted a paired T-test experiment between SED-based search and fuzzy text search using our method on all the six language pairs. We split the 50 test queries randomly into two equal, disjoint sets of 25 queries and repeated the experiment for 50 iterations to ensure smoothing in queries selection [3]. The results of the T-test experiments are shown in Table 3.

5. CONCLUSIONS

Similarities across languages of common origin and with similar writing systems have not been explored much from the IR perspective. In particular, there has been no such linguistically motivated work for languages which use writing systems of Brahmi origin. In this paper, we described a Surface Similarity based method for fuzzy string matching for performing CLIR and were able to show good improvement in performance. Our paired T-test results indicate that our retrieval scores are statistically significant. It may be noted that this method does not need any resources and, therefore, can be very useful for languages with scarce resources.

[1]Forum for Information Retrieval Evaluation.
`http://www.isical.ac.in/~clia/`
[2]`http://lucene.apache.org/`

Table 2: Comparison of CLIR results for the three different methods. BL refers the baseline method, BL+SED refers SED based method and BL+FTS respresents fuzzy text search based method

		Measures			
		MAP	P5	P10	P15
BH	BL	0.0595	0.1240	0.1160	0.1093
	BL+SED	0.0596	0.1240	0.1140	0.1093
	BL+FTS	0.1240	0.2000	0.1940	0.1880
BM	BL	0.0400	0.0480	0.0520	0.0467
	BL+SED	0.0400	0.0480	0.0520	0.0453
	BL+FTS	0.0915	0.1280	0.1080	0.1013
HB	BL	0.0511	0.0720	0.0720	0.0720
	BL+SED	0.0508	0.0720	0.0720	0.0693
	BL+FTS	0.0754	0.1400	0.1300	0.1240
HM	BL	0.1715	0.2520	0.2200	0.1893
	BL+SED	0.1741	0.2520	0.2220	0.1933
	BL+FTS	0.2007	0.2880	0.2380	0.2080
MB	BL	0.0251	0.0400	0.0320	0.0333
	BL+SED	0.0251	0.0400	0.0320	0.0333
	BL+FTS	0.0454	0.0720	0.0660	0.0707
MH	BL	0.1379	0.2680	0.2380	0.2280
	BL+SED	0.1378	0.2600	0.2340	0.2307
	BL+FTS	0.1603	0.2800	0.2820	0.2773

Table 3: Paired T-test scores comparison for BL+SED and BL+FTS

Run ID	Set1	Set2
BH	0.0376	0.0497
BM	0.0218	0.0231
HB	0.0113	0.0112
HM	0.0235	0.0152
MB	0.0112	0.0546
MH	0.0122	0.0150

6. REFERENCES

[1] T. M. Ellison and S. Kirby. Measuring language divergence by intra-lexical comparison. In *Proc. of 44th ACL*, 2006.

[2] T. Masuyama, S. Sekine, and H. Nakagawa. Automatic construction of japanese katakana variant list from large corpus. In *Proc. of 20th COLING*, 2004.

[3] M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proc. 28th ACM SIGIR*, 2005.

[4] A. K. Singh. A computational phonetic model for indian language scripts. In *Constraints on Spelling Changes: 5th Intl. Workshop on Writing Systems. Nijmegen, The Netherlands, October, 2006.*

[5] A. K. Singh, H. Surana, and K. Gali. More accurate fuzzy text search for languages using abugida scripts. In *Proc. of iNEWS07.* ACM, 2007.

[6] E. Yamamoto, M. Kishida, Y. Takenami, Y. Takeda, and K. Umemura. Dynamic programming matching for large scale information retrieval. In *Proc. of 6th Intl. workshop on Information retrieval with Asian languages.* ACL, 2003.