Improving Cross-Language Information Retrieval by Transliteration Mining and Generation

K Saravanan, Raghavendra Udupa and A Kumaran

Multilingual Systems Research
Microsoft Research India
Bangalore, India.
{v-sarak, raghavu, kumarana}@microsoft.com

Abstract. The retrieval performance of Cross-Language Retrieval (CLIR) systems is a function of the coverage of the translation lexicon used by them. Unfortunately, most translation lexicons do not provide a good coverage of proper nouns and common nouns which are often the most information-bearing terms in a query. As a consequence, many queries cannot be translated without a substantial loss of information and the retrieval performance of the CLIR system is less than satisfactory for those queries. However, proper nouns and common nouns very often appear in their transliterated forms in the target language document collection. In this work, we study two techniques that leverage this fact for addressing the problem, namely, Transliteration Mining and Transliteration Generation. The first technique attempts to mine the transliterations of out-of-vocabulary query terms from the document collection whereas the second generates the transliterations. We systematically study the effectiveness of both techniques in the context of the Hindi-English and Tamil-English ad hoc retrieval tasks at FIRE2010. The results of our study show that both techniques are effective in addressing the problem posed by out-of-vocabulary terms with Transliteration Mining technique giving better results than Transliteration Generation.

Keywords: Cross-Language Information Retrieval System, FIRE 2010, Transliteration Mining, Transliteration Generation.

1 Introduction

With the exponential increase in non-English user population on the Internet over the last two decades, Cross-Language Information Retrieval (CLIR) has gained importance both as a research discipline and as an end-user technology. The importance of this discipline is evidenced by increased number of research publications, workshops and shared tasks, focusing on various aspects of information retrieval tasks in multilingual or cross-language settings. While there has been substantial progress in the core cross-language retrieval algorithms, the retrieval performance of any CLIR system is a function of the coverage of the translation lexicon used by the system. As query terms (or their statistics) must be translated in CLIR before the retrieval of documents, translation lexicon plays a large role in determining the retrieval performance no matter what retrieval algorithm is ultimately employed by the CLIR system. When query terms cannot be translated to the target language, there could be a loss of information and consequently, a loss in retrieval performance. The loss of information is substantial when the query terms are proper nouns or common nouns which are often the information bearing terms in the query.

Unfortunately, most translation lexicons do not provide a good coverage of proper nouns (names) and it turns out that names appear often in queries and constitute the largest class of out-of-vocabulary terms in CLIR [43].[1] This is not surprising because names form an open set in a language and every day new names enter into a language. Hence, it is practically impossible to keep a translation lexicon up-to-date. Further, the same source language name can have multiple variants in the target language due to the difference in the sounds used in the two languages and also due to inflections and agglutination. Therefore, even when the translation lexicon has an equivalent for a source language name, it might not be the variant that appears in the documents relevant to the query.

In cultures where a foreign language like English is widely used (say as a second language), many common nouns in the foreign language are imported to the native language by a phenomenon called code-mixing – the foreign language words are transliterated to the native language and used instead of their equivalents in the native language. As with proper nouns, most translation lexicons do not provide a good coverage of such imported common nouns.

Proper nouns and common nouns are often the most information bearing terms of a query and can cause severe loss in retrieval performance when left untranslated. However, in many cases, proper nouns and common nouns (those which are imported from a foreign language) are found in their transliterated forms in the target language document collection. In this work we study two techniques that leverage this fact to address the problem posed by out-of-vocabulary terms in CLIR:

1. Transliteration Mining
2. Transliteration Generation.

Transliteration Mining is a novel technique that was proposed in [43] and shown to give significant improvements in retrieval performance over a language model based CLIR baseline. It employs a two-pass retrieval approach. The source language query is translated using the translation lexicon ignoring all out-of-vocabulary (OOV) terms and a first pass retrieval is done. Transliterations for the out-of-vocabulary terms are then mined from the top results using a statistical transliteration model. The source language query is now retranslated using the translation lexicon and the transliterations thus mined and a second pass retrieval is done.

[1] In fact, 60% of the topics in the 2000-2007 Cross-Language Evaluation Forum (CLEF) ad hoc retrieval tasks had at least one name and 18% of them had at least three.

Transliteration Generation has been employed in CLIR by many works including [1, 46]. It generates the transliterations of the out-of-vocabulary terms using Machine Transliteration and uses them along with the translation lexicon to translate the source language query. In this work, we employ a state of the art Machine Transliteration technique [19].

We systematically study the above mentioned techniques and evaluate their effectiveness relative to a language model based CLIR baseline. We also compare the two techniques to two oracles, which can identify for every out-of-vocabulary term, its equivalent in English topic and in the target collection, respectively. Our study is a continuation of the initial study done as part of FIRE 2010 ad hoc CLIR tasks [40].

The rest of the paper is organized as follows: In Section 2 we discuss some works that are related to our work. In Section 3 we discuss the CLIR system and two techniques for handling OOV terms. In Section 4 we discuss the experimental setup and in Section 5 we report the results.

2 Related Work

Basic CLIR systems that use translation lexicons have been studied in several works in early literature, for example [4, 18], but they suffer from the problem of OOV terms in the queries, which often are names. Leaving the OOV query terms untranslated is well recognized to have a significant negative impact on the performance of CLIR systems [8, 25, 26, 47]. Broadly, there are two distinct approaches taken to address the problem of OOV terms: The first approach is to employ a Machine Transliteration system to generate the transliteration equivalents in the target language [1, 2, 15, 17, 23, 46] , and use them for retrieval. The second approach is to enhance the translation lexicon offline, by mining the transliteration equivalents from parallel or comparable corpora [3, 10, 11, 29, 37, 43].

Several methods have been reported in CLIR literature where Machine Transliteration was employed on OOV query terms. They differ in the exact technique used for transliterating the OOV query terms to the target language [1, 14, 15, 34, 46, 51]. While most of the above works report improved retrieval performance, the improvements are modest. This is because Machine Transliteration, to be effective in CLIR, has to produce the exact string used in the document collection and not just any acceptable similar sounding string. However, we note that in the recent past there has been significant progress in Machine Transliteration as evidenced by the results of the shared task on Machine Transliteration at NEWS 2009, 2010, and 2011 workshops [20, 22, 50]. Thanks to this series of workshops, it is now possible to calibrate different Machine Transliteration techniques and use the best among them in CLIR.

Mining based approaches are used in Machine Translation to augment the translation lexicon with name transliterations. In the literature, there are many interesting corpus-based techniques for mining both translation equivalents and transliteration equivalents [5, 18, 29, 37, 39, 44, 45][2]. While many such techniques have addressed

[2] For a more detailed discussion on corpus-based mining techniques, please see [45].

the OOV problem in Machine Translation and improved the quality of translated text, none of them are effective in ad hoc retrieval tasks. This is because corpus-based mining techniques might not always be successful in finding the transliteration equivalents of the specific names and common nouns used in the topics of an ad hoc retrieval task. Therefore, the best place to look for transliterations of OOV terms of a query are the top results from the target collection itself for the query as hypothesized in [43]. In this approach, the source language query is translated using the translation lexicon ignoring all out-of-vocabulary (OOV) terms and a first pass retrieval is done. Transliterations for the out-of-vocabulary terms are then mined from the top results using a statistical transliteration model. This is the approach we also take in this study.

We at Microsoft Research India (MSR India) [28] fielded a CLIR system without addressing the OOV query terms in the CLEF 2007 [41] campaign for the Hindi-English track [13]. In FIRE 2008 campaign, we fielded a CLIR system that employed Transliteration Mining in Hindi-English track [42]. In 2010, FIRE organized several ad hoc monolingual and cross-language retrieval tracks, and we fielded a CLIR system that used both Transliteration Mining and Transliteration Generation – in the cross-language Hindi-English and Tamil-English ad hoc retrieval tracks [40].

3 Retrieval System

In this section, we outline the various components of our CLIR system that participated in FIRE2010.

3.1 Monolingual Retrieval System

Our monolingual retrieval system is based on the well-known Language Modeling framework for Information Retrieval [35, 49]. In this framework, queries as well the documents are viewed as probability distributions. The similarity of a query (*q*) with a document (*d*) is measured in terms of the likelihood of the query under the document language model (or equivalently, as the Kullback-Leibler divergence of query and document unigram language models).

$$Score(q,d) = \sum_{w} p(w \mid q) \log p(w \mid d) \tag{1}$$

where, *w* is the term in the lexicon. For a detailed description and discussion of the Language Modeling framework, please see [35, 48, 49]. We smooth the document language model by interpolating with a corpus language model:

$$p_{sm}(w \mid d) = (1-\alpha) p_{mle}(w \mid d) + \alpha p(w \mid C) \tag{2}$$

where $0 \leq \alpha \leq 1$.

3.2 Cross-Language Retrieval System

We translate the query distribution in the source language (q_s) to the target language using a probabilistic translation lexicon:

$$p(w_t \mid q_s) = \sum_{w_s} p(w_s \mid q_s) p(w_t \mid w_s) \tag{3}$$

where w_s is a source language term and w_t is a target language term. Note that the target language translation (q_t) of the query need not have a surface realization. Nevertheless, the similarity of the translated query (q_t) with a document (d_t) is measured in terms of the Kullback-Leibler divergence of the query and the document language models:

$$\begin{aligned} Score(q_s, d_t) &= \sum_{w_t} p(w_t \mid q_t) \log p(w_t \mid d_t) \\ &= \sum_{w_t, w_s} p(w_s \mid q_s) p(w_t \mid w_s) \log p(w_t \mid d_t) \end{aligned} \tag{4}$$

3.3 Handling Out-of-Vocabulary terms

Like any cross-language system that makes use of a translation lexicon, we too faced the problem of OOV query terms. As we observed earlier, many of the OOV terms are names that can be transliterated to the target language and some are imported common nouns. To handle these OOV terms, we used two different techniques, (i) Transliteration Mining and (ii) Transliteration Generation. The details of the above two techniques are given in subsequent sections.

3.4 Transliteration Mining

The mining algorithm issues the translated query minus OOV terms to the target language information retrieval system and mines transliterations of the OOV terms from the top results of the first-pass retrieval. Hence, in the first pass, each query-result pair is viewed as a “comparable” document pair, assuming that the retrieval brought in a reasonably good quality results set based on the translated query without the OOV terms. The algorithm hypothesizes a match between an OOV query term and a document term in the “comparable” document pair and employs a transliteration similarity model (Section 3.4.1) to decide whether the document term is a transliteration of the query term [43, 45]. Transliterations mined in this manner are then used to retranslate the query and issued again, for the final retrieval.

For each topic, we considered top-100 documents returned by the cross-language retrieval system for the purpose of mining. We refer to [43] for details of the mining technique.

3.4.1 Transliteration Similarity Model

Our transliteration similarity model is an extension of W-HMM word alignment model presented in [12], which had been shown to perform well on Transliteration Mining tasks [43, 44, 45]. It is a character-level hidden alignment model that makes use of a richer local context in both the transition and emission models compared to the classic HMM model [25]. The transition probability depends on both the jump width and the previous source character as in the W-HMM model. The emission probability depends on the current source character and the previous target character unlike the W-HMM model. The transition and emission models are not affected by data scarcity unlike Machine Translation as the character lexicon of a language is typically several orders smaller than its word lexicon. Instead of using any single alignment of characters in the pair (w_S, w_T), we marginalize over all possible alignments:

$$p(t_1^m \mid s_1^n) = \sum_A \prod_{j=1}^{m} p(a_j \mid a_{j-1}, s_{a_{j-1}}) p(t_j \mid s_{a_j}, t_{j-1}) \quad (5)$$

Here, t_j (respectively, s_i) denotes the j^{th} (respectively, i^{th}) character in target word w_T (respectively, source word w_S) and $A \equiv a_1^m$ is the hidden alignment between w_T and w_S where t_j is aligned to s_{a_j}, $j = 1,...,m$. We estimate the parameters of the model by learning over a training set of transliteration pairs. We use the EM algorithm to iteratively estimate the model parameters. The transliteration similarity score of a pair (w_S, w_T) is $\log p(w_T \mid w_S)$ appropriately transformed.

3.5 Transliteration Generation

We experimented with two different techniques of generating transliterations in a target language – Direct and Compositional. In direct transliteration, the OOV terms are directly transliterated using a Machine Transliteration system trained on source-target language parallel names corpora, as detailed in Section 3.5.1.

In compositional transliteration, we use a two-stage system as outlined in Section 3.5.2, for generating transliterations of a given OOV term in source language into the target language, by transitioning through an intermediate language.

3.5.1 Direct Transliteration Generation

Systematic comparison of the various transliteration systems in the NEWS-2009 workshop [22] showed conclusively that orthography based discriminative models like Conditional Random Fields [21] performed best in a language-neutral manner. Hence, we decided to adopt a conditional random fields-based approach using purely orthographic features. In addition, we introduced a word origin detection module to identify specifically Indian origin names. Use of such classifiers allowed us to train a specific CRF-based transliteration engine for Indian origin names, and thus producing better quality transliterations. All other names are transliterated through an engine that is trained on non-Indian origin names. Such a system was used for generating transliterations of OOV terms between source and target languages.

For word origin detection, we manually classified 3000 words from the training set (detailed in Section 4.3.1) into words of Indic origin and Western origin. Two trigram language models were built, one for the Indic origin names and another for Western origin names, to help classify all the name pairs in the training set as Indic or Western names. Manual verification showed that this method about 97% accurate, yielding good quality data that is used for training two distinct CRF-based modules for transliterating Indic and Western names.

Conditional Random Fields are undirected graphical models used for labeling sequential data [21]. Under this model, the conditional probability distribution of the target string given the source string is given by:

$$p(Y / X; \lambda) = \frac{1}{Z(X)} \cdot e^{\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k (Y_{t-1}, Y_t, X, t)} \tag{6}$$

where,

X = source string
Y = target string
T = length of source string
K = number of features
λ_k *= feature weights*
Z(X) = normalization constant
f = 1 if the feature is active, 0 otherwise

We used CRF++, an open source implementation of CRF for training and further transliterating the names (top-n most probable sequences) [7]. We used the alignment model developed by [43] to get the character level alignments for the parallel names in the training corpora. Under this alignment, each character in the source word is aligned to zero or more characters in the corresponding target word. We trained a transliteration engine, based on a rich feature set generated from this character-aligned data; the feature set includes aligned characters in each direction within a small distance (typically, 2) and source and target bigrams and trigrams.

3.5.2 Compositional Transliteration Generation

Compositional transliterations systems combine multiple direct transliterations systems serially to produce transliterations between source language to target language [16, 19]. Specifically, we assume that parallel names corpora are available between the language pair, X and Y, and the language pair, Y and Z; we train two CRF based transliteration systems (as outlined in the earlier section), between the language X and Y, and Y and Z. We provide every name in the test set (in language X) as an input to the X→Y transliteration system, take the top-10 candidate output strings (in language Y) and provide each as an input to the Y→Z system. The output of the Y→Z system for the top-10 candidate strings (in language Z) were merged and re-ranked by their probability scores. Finally, the top-10 of the merged output was taken as the final output of the compositional transliteration system.

4 Data for Experimental Setup

In this section, we specify all the data used in our experiments.

4.1 FIRE Data

The English document collection provided by FIRE2010 was used in all our runs [9]. The English document collection consists of ~124,000 news articles from "The Telegraph India" from 2004-07. All the English documents were stemmed using the Porter stemmer [36]. We ignored the stop words in the documents as well as the queries. We did not stem the query terms, due to the non-availability of good stemmers in these languages. We plan to experiment with language-neutral stemming techniques for Indian languages in our future work [27].

Totally 50 topics were provided in each of the languages, each topic having a title (T), description (D) and narrative (N), successively expanding the scope of the query. Table 1 shows a typical topic in Hindi, and the TDN components of the topic, for which relevant English documents are to be retrieved from the aforementioned English news corpus.

Table 1. A FIRE2010 Topic in Hindi.

Type	Topic
Title	गुटखा मालिकों का अन्डरवर्ल्ड के साथ उलझाव
Description	प्रसिद्ध गुटखा कम्पनी (माणिकचन्द और गोवा)के साथ दाऊद इब्राहिम के सम्बन्ध
Narration	प्रासंगिक प्रलेख में माणिकचन्द गुटखा और गोवा गुटखा मालिकों का अन्डरवर्ल्ड डॅन दाऊद इब्राहिम के साथ सम्बन्ध, से सम्बन्धित सूचनाएँ यहाँ होनी चाहिये। अन्य कम्पनियों के साथ दाऊद इब्राहिम के सम्बन्ध यहाँ अप्रासंगिक हैं।

Table 2. A FIRE2010 Topic in English.

Type	Topic
Title	Links between Gutkha manufacturers and the underworld.
Description	Links between the Goa and Manikchand Gutkha manufacturing companies and Dawood Ibrahim.
Narration	A relevant document should contain information about the links between the owners of the Manikchand Gutkha and Goa Gutkha companies and Dawood Ibrahim, the gangster. Information about links between Dawood Ibrahim and other companies is not relevant.

It should be noted that FIRE has also released a set of 50 English (i.e., target language) topics, equivalent to each of the source language topics. The purpose of such topics is to have monolingual (in target language) runs that may provide an upper bound on the retrieval performance of the CLIR runs. A typical English topic (equivalent of the one in the table above) is given in Table 2.

4.2 Bilingual Dictionaries for CLIR

For both the Hindi-English and Tamil-English cross-language retrieval tasks, statistical dictionaries were used; these statistical dictionaries were generated by training statistical word alignment models on Hindi-English parallel corpora (~100 K parallel sentences) and Tamil-English parallel corpora (~50 K parallel sentences) using the GIZA++ tool [32]. We used 5 iterations of IBM Model 1 and 5 iterations of HMM [32]. In the Hindi-English language pair, the training ultimately yielded a statistical dictionary consisting of ~59 K Hindi words and ~63 K English words. In the Tamil-English language pair, the training yielded a statistical dictionary consisting of ~107 K Tamil words and ~45 K English words. We used only top 4 translations for every source word, an empirically determined limit to avoid generation of noisy terms in the query translations.

4.3 Training Data for Transliteration Generation

4.3.1 Training Direct Transliteration Systems

The direct transliteration systems were trained with about 15 K parallel names in Hindi and English and Tamil and English. As reported in [16], the quality of a Machine Transliteration system trained with 15 K corpora is similar to that of a system trained with much larger training data, and hence we used about 15 K parallel names for training a CRF-based transliteration generation system, as described in Section 3.5.1.

4.3.2 Training Compositional Transliteration Systems

The compositional transliteration systems chains two distinct transliteration systems, as described in Section 3.5.2, each trained with about 15 K of appropriate parallel names corpora [16, 19]. In our case, we used Kannada, an Indian language of Dravidian family, as the intermediate language, and trained two separate systems: one between Hindi and Kannada, and another between Kannada and English. Kannada was chosen as the intermediate language as it has a near superset of phoneme inventory of Hindi and English, and hence captures the phonetic essence of the source name to reproduce in the target language. The compositional transliteration technique was used only for Hindi-English cross-language runs. We used top 5 results from transliteration generation for query translation.

4.4 Training Data for Transliteration Mining

We trained Hindi-English and Tamil-English transliteration similarity models on 16 K parallel single word names in Hindi-English and Tamil-English language pairs respectively, and ran 15 iterations of Expectation Maximization training.

5 Results and Analysis

In this section, we present our experimental results and also an analysis of the results.

5.1 Metrics for Measuring Performance

We use Mean Average Precision (MAP) as the measure for the topic set, Average Precision (AP) for individual topics and Precision at top-10 (P@10).

5.2 An illustrative analysis of the impact of different techniques

In this section we show one example Hindi topic and discuss how various approaches affected the retrieval performance. Note that the results of our CLIR experiments, as presented in Tables 7 & 8, indicate that the approaches generally help in improving the CLIR performance.

Consider the topic number 112 in Hindi shown in table 3. The OOV terms in the Hindi topic are shown in bold, and those OOV terms that are transliteratable are underlined; hence Transliteration Mining and Transliteration Generation can potentially help the retrieval performance, by providing equivalents in the target language for these words.

Table 3. Hindi Topic No. 112.

Type	Topic
Title	गुटखा मालिकों का **अन्डरवर्ल्ड** के साथ **उलझाव**
Description	प्रसिद्ध गुटखा कम्पनी (**माणिकचन्द** और गोवा)के साथ दाऊद इब्राहिम के सम्बन्ध
Narration	प्रासंगिक **प्रलेख** में **माणिकचन्द** गुटखा और गोवा गुटखा मालिकों का **अन्डरवर्ल्ड** **डॅन** दाऊद इब्राहिम के साथ सम्बन्ध, से सम्बन्धित सूचनाएँ यहाँ होनी चाहिये। अन्य कम्पनियों के साथ दाऊद इब्राहिम के सम्बन्ध यहाँ अप्रासंगिक हैं।

The Hindi topic has five OOV terms (namely, ‘**अन्डरवर्ल्ड**’, ‘**उलझाव**’, ‘**माणिकचन्द**’, ‘**प्रलेख**’ and ‘**डॅन**’), out of which two of them (‘**अन्डरवर्ल्ड**’, a common noun imported to Hindi from English by transliterating the word ‘*underworld*’ and ‘**माणिकचन्द**’, a proper noun ‘*manickchand*’) are terms that may occur in the transliterated form in the target language document collection. Transliteration Mining was able to identify the valid English equivalents for both of these two terms from the top results of the first pass retrieval (‘*underworld’* for ‘अन्डरवर्ल्ड’ and ‘*manikchand’* for ‘माणिकचन्द’), whereas generation produced only one English equivalent (‘*manikchand’* for ‘माणिकचन्द’) correctly.

Table 4 shows the generated and mined equivalents for the OOV terms of the Hindi topic 112 (shown in Table 3). The OOV terms that are underlined are those that occur in their transliterated form in the target corpus, and their valid English equivalents by Transliteration Generation or Transliteration Mining are shown in bold.

Table 4. OOV terms in Hindi topic 112, and their
top-5 generated transliterations (Direct and Compositional) and mined English equivalents

OOV in Topic	Generation (Direct)	Generation (Compositional)	Mining
अन्डरवर्ल्ड	andrverld, anderverld, andrverd, anderverd, andrvorld	anderverld, onderverld, enderverld, inderverld, xanderverld	**underworlds, underworld**
उलझाव	uljhav, ulwav, ulzav, ulqav, ulav	ullav, ulwav, uljav, ullau, ulzav	-
माणिकचन्द	manikchanda, **manikchand**, maanikchanda, maanikchand, manikanda	**manikchand**, manikchandh, manikchande, manikchandr, maanikchand	**manikchand, manickchand**
प्रलेख	pralekh, pralekha, prlekh, pralaekh, pralakh	pralekh, pralekha, pralekh, pralekh, pralek	palekar
डॅन	dann, dan, den,denn, danne	don, den, dan, dn, dian	

Note each of the OOV terms were handled slightly differently by the two competing transliteration techniques: The English equivalent for the Hindi OOV term ‘अन्डरवर्ल्ड’ (code-mixed Hindi word for the English word, ‘*underworld’*) was not generated correctly by Transliteration Generation techniques, but its two equivalents were mined correctly by Transliteration Mining. The English equivalent for the Hindi OOV term ‘माणिकचन्द’ (a transliteration for the proper noun, ‘*manikchand’*) was generated correctly by both the generation techniques, but Transliteration Mining was able to mine multiple variants of the name from the target corpus.

The two Hindi OOV terms, namely ‘उलझाव’ and ‘प्रलेख’, are not transliteratable (that is, they are proper Hindi words that were not translated by our query translation engine, due to the lack of coverage, and its transliterated form is unlikely to be in the target corpus). For the term ‘उलझाव’, the Transliteration Generation techniques produced some English strings (which clearly will not be found in the target English corpora), and Transliteration Mining also could not mine any equivalents. However, for the term ‘प्रलेख’, Transliteration Mining did find a near phonetic equivalent ‘*palekar’* which occurs in the target corpus but semantically unrelated to the source word

‘**प्रलेख**’; The generated equivalents for the OOV term ‘**प्रलेख**’ may have had relatively small negative effect on retrieval performance, as they are noisy terms.

We observe that both Transliteration Generation and Transliteration Mining introduced some noise words as well along with the correct transliterations. However, it is important to note that the positive effect of handling OOV terms correctly outweighs the negative effect of noisy terms which are in general uncorrelated with the query terms. As can be observed in Figure 1, the overall retrieval performance of topic 112 is significantly improved when either of the techniques is employed.

5.3 Performance of various configurations of Integrated CLIR System

As shown in Table 1, each of the 50 topics in Hindi and Tamil has a title (T), description (D) and narrative (N), successively expanding the scope of the query. We ran our experiments taking progressively each of (title), (title and description), and (title, description and narrative), calibrating the cross-language retrieval performance at each stage, to explore whether expanding the query adds useful information for retrieval or just noise. Table 5 shows the notation used in our description of various configurations to interpret the results presented in Tables 6, 7 and 8.

Table 5. Notations used.

T	Title
TD	Title and Description
TDN	Title, Description and Narration
M	Transliteration Mining
G_D	Transliteration Generation – Direct
G_T	Transliteration Generation – Compositional

Tables 6, 7 and 8 show the MAP and precision@10 of our monolingual as well as cross-language official runs submitted to FIRE 2010 shared task. The format of the run ids in the results table is ‘Source-Target-Query-Technique’, where ‘Query indicates the type of the query, and is one of {T, TD, TDN} and ‘Technique’ indicates the technique and from the set {M, G_D, G_T, M+G_D, M+G_T}. The ‘+’ refers to the combination of more than one approach. The symbols double star (**) and single star (*) indicate statistically significant differences with 95% and 90% confidence respectively according to the paired t-test over the baseline. The best results achieved are highlighted in bold.

5.4 Monolingual English Retrieval

We submitted 3 official runs for the English monolingual track, as shown in the Table 6. For these runs, the English topics provided by the FIRE 2010 organizers were used.

Table 6. English Monolingual Retrieval Performance (Official submissions for the FIRE 2010 Shared Task).

Run	MAP	P@10
English-English-T	0.3653	0.344
English-English-TD	0.4571	0.406
English-English-TDN	**0.5133**	**0.462**

With the full topic (TDN), our monolingual IR system achieved a MAP score of 0.5133. Generally this performance is thought to be the upper bound for cross-language performance, presented in Tables 7 and 8.

5.5 Hindi-English Cross-Language Retrieval

We submitted totally 18 official run results on Hindi-English cross-language track, as shown in Table 7.

Table 7. Hindi-English Cross-Language Retrieval Performance (Official submissions for the FIRE 2010 Shared Task).

Run	MAP	P@10
Hindi-English-T	0.2931	0.26
Hindi-English-T[GD]	0.3168**	0.282
Hindi-English-T[GT]	0.3140**	0.276
Hindi-English-T[M]	**0.3390****	**0.304**
Hindi-English-T[M+GD]	0.3388**	0.302
Hindi-English-T[M+GT]	0.3388**	0.302
Hindi-English-TD	0.4042	0.356
Hindi-English-TD[GD]	0.4336**	0.386
Hindi-English-TD[GT]	0.4369**	0.382
Hindi-English-TD[M]	0.4376**	**0.388**
Hindi-English-TD[M+GD]	**0.4378****	0.386
Hindi-English-TD[M+GT]	0.4375**	0.386
Hindi-English-TDN	0.4748	0.424
Hindi-English-TDN[GD]	0.4942**	0.434
Hindi-English-TDN[GT]	0.4970**	0.438
Hindi-English-TDN[M]	**0.4977****	0.442
Hindi-English-TDN[M+GD]	0.4971**	**0.444**
Hindi-English-TDN[M+GT]	0.4965**	**0.444**

The first run under each of the ‘T’, ‘TD’ and ‘TDN’ sections in Table 7 present the results of the runs of our baseline CLIR system without handling the OOV terms, and hence provide a baseline for measuring the improvement in retrieval performance due to Transliteration Generation or Transliteration Mining, provided subsequently. From the results, we observe that the usage of all of the components of the topic, namely T, D and N, produced the best retrieval performance. The basic Hindi-English cross-language run ‘Hindi-English-TDN’ (without Transliteration Generation or Transliteration Mining), achieved the MAP score 0.4748, and our best cross-language run ‘Hindi-English-TDN[M]’ with Transliteration Mining achieved a MAP score of 0.4977. We observe similar trends in the other runs that use only the title, or title and description sections of the topics. It should be noted that our basic TDN run achieves 92% of the monolingual performance, and the cross-language TDN run enhanced with Transliteration Mining, 97% of the monolingual retrieval performance.

In practice, user queries are more likely to be the topics restricted to ‘T’. We note that for the ‘T’ runs, Transliteration Mining gives superior retrieval performance than Transliteration Generation.

5.6 Tamil-English Cross-Language Retrieval

We submitted totally 12 official Tamil-English cross-language runs, as shown in Table 8. As with the Hindi-English runs, the first run under each of the ‘T’, ‘TD’ and ‘TDN’ sections in Table 8 present the results of the runs without handling the OOV terms, and is a baseline.

Table 8. Tamil-English Cross-Language Retrieval Performance (Official submissions for the FIRE 2010 Shared Task).

Run	**MAP**	**P@10**
Tamil-English-T	0.2710	0.258
Tamil-English-T[GD]	**0.2891***	**0.268**
Tamil-English-T[M]	0.2815**	0.258
Tamil-English-T[M+GD]	0.2816*	0.268
Tamil-English-TD	0.3439	0.346
Tamil-English-TD[GD]	0.3548*	0.35
Tamil-English-TD[M]	**0.3621****	0.346
Tamil-English-TD[M+GD]	0.3617**	**0.362**
Tamil-English-TDN	0.3912	0.368
Tamil-English-TDN[GD]	0.4068**	0.378
Tamil-English-TDN[M]	**0.4145****	0.368
Tamil-English-TDN[M+GD]	0.4139**	**0.394**

From the results presented in Table 8, we observe that the usage of all of the components of the topic, namely T, D and N, produced the best retrieval performance. The basic Tamil-English cross-language run ‘Tamil-English-TDN’ achieved the MAP score 0.3912, and our best cross-language run ‘Tamil-English-TDN[M]’ with Transliteration Mining achieved a MAP score of 0.4145. We observe, in general, similar trends in the other runs that use only the title, or title and description sections of the topics. While the cross-language performance of Tamil-English achieves ~81% of our monolingual English retrieval performance, we observe that this is not as high as the

Hindi-English retrieval, perhaps due to the highly agglutinative nature of Tamil as we explain in section 5.7.2.

Given that Transliteration Mining performed generally better than Transliteration Generation, we take a deeper look at Transliteration Mining in the next section.

5.7 Mining OOV terms and its effect on CLIR performance

In this section, we analyze the volume of the OOV terms in FIRE topics, and to what extent they are handled by Transliteration Mining, which clearly emerged as the better technique for addressing the OOV problem. Also, we show the effect of handling the OOVs on the cross-language retrieval performance, for both the Hindi-English and Tamil-English CLIR runs.

5.7.1 Profile of OOV terms in Hindi-English Cross-lingual Task

Table 9 gives the profile of OOV terms. We see that there are a large number of OOV terms in all three query configurations and a good number of queries are affected. We also see that majority of the OOV terms are transliteratable, i.e. they are either names or imported common nouns. Further, Transliteration Mining is able to mine at least one correct transliteration for most of the transliteratable OOV terms.

Table 9. Profile of OOV terms in Hindi-English CLIR.

Type	All OOV terms		Transliteratable OOV terms		Transliteratable OOV terms handled correctly	
	No. of Terms	No. of Topics	No. of Terms	No. of Topics	No. of Terms	No. of Topics
T	15	13	11	11	11	11
TD	35	24	23	17	21	15
TDN	73	50	31	19	24	17

Table 10. Performance Improvements in Hindi-English CLIR

Type	MAP as % of monolingual	MAP improvement in % over the baseline	
		All topics	Only topics with OOV
T	92.8	15.7	60.7
TD	95.7	8.3	14.1
TDN	97.0	4.8	4.8

Table 10 shows the performance improvements that Transliteration Mining brings relative to the baseline. The percentage of MAP score improvement over the monolingual performance is shown in column 2. The percentage improvements over the baseline CLIR system are measured in two contexts: with all topics in the FIRE cross-lingual task, and with only those topics that have at least one OOV in them. These two are shown in columns 3 and 4 respectively.

In the TD configuration, the 50 Hindi topics contained 35 distinct OOV terms that appeared in 24 topics. Out of these 35 terms, 23 were proper or common nouns and appeared in 17 topics. Transliteration Mining produced at least one transliteration equivalent for 21 of these OOV terms (91.3%), which appeared in totally15 topics. As shown in Table 7 & 10, handling these OOV terms resulted in MAP score improvement of 8.3% when considering all 50 topics and 14.1% when considering only those 24 topics that have at least one OOV term, over the baseline CLIR system. Similarly, in the T configuration, Transliteration Mining produced at least one transliteration equivalent for all 11 transliteratable terms (that is, 100%) that appeared in 11 topics. As a consequence, the MAP improved by 15.7% when considering all 50 topics and by 60.7% when considering only those 13 topics that had at least one OOV term. This highlights the significance of our method in practice where most queries are short.

5.7.2 Profile of OOV terms in Tamil-English Cross-lingual Task

Table 11 gives the profile of OOV query terms and Table 12 shows the performance improvements for our Tamil-English CLIR system.

Table 11. Profile of OOV terms in Tamil-English CLIR.

Type	All OOV terms		Transliteratable OOV terms		Transliteratable OOV terms handled correctly	
	No. of Terms	No. of Topics	No. of Terms	No. of Topics	No. of Terms	No. of Topics
T	24	19	13	13	5	5
TD	58	33	29	21	15	12
TDN	129	45	47	28	24	17

As shown in Table 11, in the TDN configuration, the 50 Tamil topics contained 129 unique OOV terms that appeared in 45 topics, out of which 47 (in 28 topics) were proper or common nouns. Transliteration Mining produced at least one transliteration equivalent for 24 of these OOV terms (51.06%) that appeared in 17 topics. As shown in Table 8 and 12, handling these OOV's resulted in MAP score improvement by 6% when all 50 topics are considered and 6.5% when considering only the 45 topics that had at least one OOV term.

As we observed in Hindi-English task, short queries benefit most (13.6%) from Transliteration Mining in Tamil-English task also. However, this is relatively low when compare to Hindi-English task (60.7%). This is due to the reason that in Hindi-English task 100% of transliteratable OOV terms were handled correctly whereas it is only 38.46% in Tamil-English task. Also, note that the performance of our Tamil-

English CLIR system with Transliteration Mining is ~81% of the monolingual performance in TDN configuration.

Table 12. Performance Improvements in Tamil-English CLIR

Type	MAP as % of monolingual	MAP improvement in % over CLIR baseline	
		All topics	Only topics with OOV
T	77.1	3.9	13.6
TD	79.2	5.3	8.7
TDN	80.8	6.0	6.5

Tamil poses specific challenges compared to Hindi: First, the transliteratable terms mentioned in the fourth column of the Table 11 excludes some terms whose equivalents are multiword expression in English; mining such multiword transliteration equivalents is beyond the scope of our work and hence they were not handled. Second, 26 out of the 47 terms that are transliteratable were inflected or agglutinated. While Transliteration Mining algorithm could mine some of them, many could not be at the threshold used by the transliteration similarity model. By relaxing the threshold we could mine more such terms, but that introduced many more noise terms, affecting the overall retrieval performance. We believe that the use of a good stemmer for inflectional languages like Tamil may help our Transliteration Mining algorithm and thereby the cross-language retrieval performance.

5.8 Mining OOV terms and its effect on individual topic performance

In this section, we discuss the effect of Transliteration Mining on the retrieval performance for individual topics of the FIRE 2010 shared task. Figure 1 shows the difference in the Average Precision – topic-wise – between the baseline CLIR system and the one that employs Transliteration Mining. We see that many topics benefitted from Transliteration Mining; for example, in the Hindi-English language pair, in T configuration, 8 topics benefitted substantially (with improvement of ≥ 0.2 in AP) whereas only 2 topics were negatively impacted (a drop of ≥ 0.2 in AP). Similar trends could be seen for all configurations, in both Hindi-English and Tamil-English language pairs.

In order to observe the impact of Transliteration Mining on individual topics, let us consider some topics in the Hindi-English and Tamil-English test collections (in TDN configuration), specifically those topics which are affected most. Such topics are shown in the Table 13. The OOV terms of these topics are shown in bold, and those OOV terms that are transliteratable are underlined. Subsequently, we show how the Transliteration Mining handled each of such OOV terms, and present the resulting change on the retrieval performance over the baseline CLIR. Table 14 shows the

OOV terms of the above topics and the outputs of Transliteration Mining. The valid English equivalents mined are shown in bold.

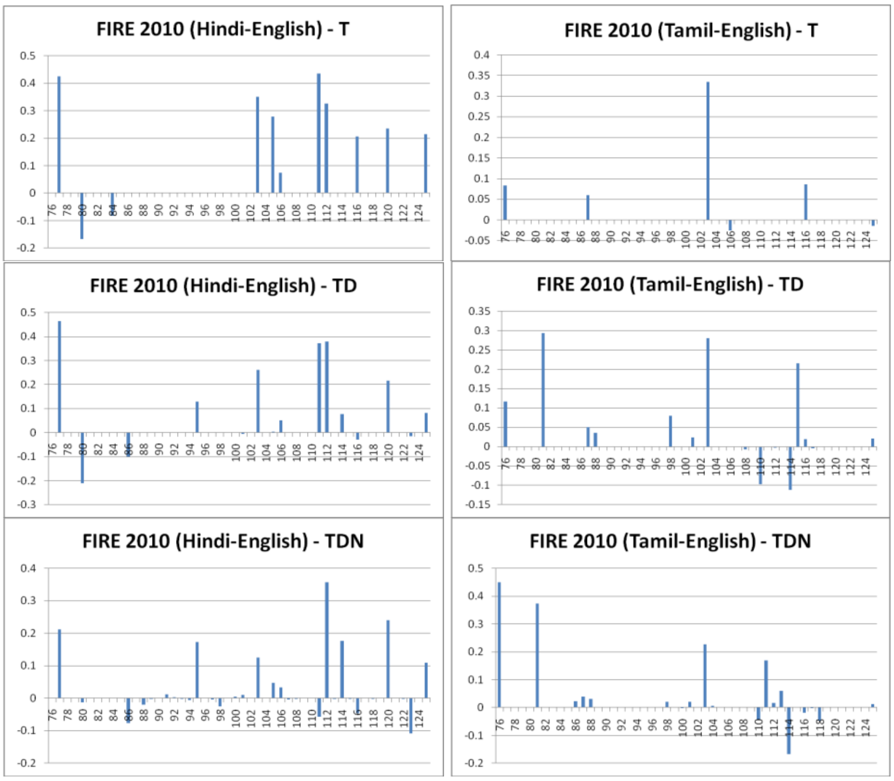

Fig. 1. Differences in Average Precision between the baseline and CLIR with Transliteration Mining

In Hindi-English task, the AP of topic 112 was increased by +0.36 in the TDN setup. This topic has five Hindi OOV terms, out of which two of them (‘**अन्डरवर्ल्ड**’ and ‘**माणिकचन्द**’) are transliteratable. Transliteration Mining was able to produce all the valid English equivalents from the top results, but it also mined a noisy term for non-transliteratable OOV term ‘**प्रलेख**’. Still, the overall effect on the retrieval performance was positive. On the other hand, topic 123 has two OOV terms, namely ‘**फलस्तीनी**’and ‘**प्रलेख**’, out of which only one (‘**फलस्तीनी**’) is transliteratable; Transliteration Mining was able to produce two equivalents. However, as in topic 112, Transliteration Mining produced a noisy term for the non transliteratable OOV ‘**प्रलेख**’; the overall effect on the retrieval performance was negative (-0.11). When we investigated this topic further, we found that the relevant documents had the term ‘*palestinian’* which was being stemmed as ‘*palestinian’* where as the mined transliterations, ‘*palestines’* and ‘*palestine’*, were both stemmed as ‘*palestin’*. Had the stemmer produced the same stems for ‘*palestinian’*, ‘*palestines’* and ‘*palestine’*, the retrieval performance would have gone up for this topic.

Table 13. Topics affected most by Transliteration Mining.

Topic ID	Topic
Hindi 112	(T) गुटखा मालिकों का **अन्डरवर्ल्ड** के साथ **उलझाव** (D) प्रसिद्ध गुटखा कम्पनी (**माणिकचन्द** और गोवा)के साथ दाऊद इब्राहिम के सम्बन्ध (N) प्रासंगिक **प्रलेख** में **माणिकचन्द** गुटखा और गोवा गुटखा मालिकों का **अन्डरवर्ल्ड** **डॅन** दाऊद इब्राहिम के साथ सम्बन्ध, से सम्बन्धित सूचनाएँ यहाँ होनी चाहिये। अन्य कम्पनियों के साथ दाऊद इब्राहिम के सम्बन्ध यहाँ अप्रासंगिक हैं।
Hindi 123	(T) यासर अराफात की मृत्यु (D) **फलस्तीनी** नेता यासर अराफात की मृत्यु (N) प्रासंगिक **प्रलेख** में फलस्तीनी नेता यासर अराफात की मृत्यु से सम्बन्धित सूचनाएँ होनी चाहिये। फलस्तीनी नेता की मृत्यु से फैली राजनीतिक अशांति से सम्बन्धित सूचनाएँ यहाँ अप्रासंगिक हैं
Tamil 76	(T) குஜ்ஜார், **மீனாஸ்** இடையே மோதல் (D) பழங்குடியினர் பட்டியலில் **குஜ்ஜாரை** **இணைத்ததற்கு** **மீனாஸ்** தலைவர்களின் எதிர்ப்பு. (N) பழங்குடியினர் பட்டியலில் **இணைக்க** வேண்டும் என **குஜ்ஜார்களின்** கிளர்ச்சி. **மீனாஸ்** தலைவர்கள் இதற்கு கடும் எதிர்ப்பு. **மீனாஸ்** தலைவர்கள் எதிர்ப்பதற்கான பின்னணி காரணங்கள் என்ன? இவ்விரு **பிரிவினரிடையேயான** போராட்டத்திற்கு **மூலக்காரணம்** பற்றிய செய்திகள் இந்த ஆவணத்தில் இடம்பெறலாம்.
Tamil 114	(T) பாதுகாப்புத் துறையின் ஆயுத ஊழல் விசாரணை (D) ஜார்ஜ் **பெர்ணான்டர்ஸ்**, டெனில் இடையேயான ஆயுத ஒப்பந்தம், இந்த **முறைக்கேடிற்கு** பிரணாப் முகர்ஜியின் **விசாரணைத்** தேவை என்ற **கோரிக்கைப்** பற்றிய செய்திகள் இந்த ஆவணத்தில் இடம்பெறலாம். (N) முன்னாள் அமைச்சர் ஜார்ஜ் **பெர்ணான்டர்ஸ்**, தென்னாப்பிரிக்க நிறுவனமான **டெனிலுடனான** ஆயுத ஒப்பந்தம். இந்த ஒப்பந்தம் குறித்த பாதுகாப்பு அமைச்சர் பிரணாப் முகர்ஜியின் **விசாரணைப்** பற்றிய தகவல்கள் இந்த ஆவணத்தில் இடம்பெறலாம்.

In the Tamil-English task, the topic 76 had 7 OOV terms, out of which only 3 are transliteratable. Transliteration Mining produced the equivalent for one of them (which occurred 4 times in the topic) and as a consequence, the AP increased by 0.45. Topic 114 has 6 OOV terms, out of which only two are transliteratable. Transliteration Mining produced a valid transliteration ‘*fernandess’* for the transliteratable OOV term ‘**பெர்ணான்டர்ஸ்**’; but this mined term is different from that in the equivalent English topic, which is ‘*fernandez’* and AP went down by-0.17).

Table 14. Topics with their OOV terms and the impact on retrieval performance.

Topic ID	Hindi/Tamil OOV	Mined English words	Change in AP
Hindi 112	अन्डरवर्ल्ड	**underworlds, underworld**	+0.36
	उलझाव	-	
	माणिकचन्द	**manikchand, manick-chand**	
	प्रलेख	palekar	
	डॅान	-	
Hindi 123	फलस्तीनी	**palestines, palestine**	-0.11
	प्रलेख	palekar	
Tamil 76	மீனாஸ்	menace**, meenas**	+0.45
	குஜ்ஜாரை	-	
	இணைத்ததற்கு	-	
	இணைக்க	-	
	குஜ்ஜார்களின்	-	
	பிரிவினரிடையேயான	-	
	மூலக்காரணம்	-	
Tamil 114	பெர்ணான்டர்ஸ்	**fernandess**	-0.17
	முறைக்கேடிற்கு	-	
	விசாரணைத்	-	
	கோரிக்கைப்	-	
	டெனிலுடனான	-	
	விசாரணைப்	-	

5.9 Hybrid approach: Mining with Transliteration Generation

In addition to generation and mining of transliteration equivalents we conducted a few experiments that employed a combination of both techniques. In this technique, we first mine the transliteration equivalents using Transliteration Mining and employ Transliteration Generation for those OOV terms for which Transliteration Mining produced no results. In Table 7 and 8, the run ids with ‘M+G_D’ and ‘M+G_T’, refer that they are combination of mining and generation. We observe that the hybrid approach did not produce significantly better results than Transliteration Mining in both Hindi-English and Tamil-English.

5.10 Comparison of Transliteration Mining against Oracles

Finally, in this section, we compare the performance of our best performing configuration –Transliteration Mining – against two oracular systems, which can identify the right transliterations from equivalent English topic and relevant documents from the target collection, respectively, and thus can indicate an upper bound for the cross-language retrieval performance. We devised two oracular CLIR systems, as described below. The first oracular system identified the correct transliterations for OOV terms from the equivalent English topic (which was provided as a part of the FIRE 2010 test collection). The second oracular system identified the correct transliteration equivalents for the OOV terms from the relevant documents for that topic (provided as a part of the FIRE 2010 test collection. Note that in both the above oracles, we used the same statistical dictionaries used in our previous experiments. The MAP figures for the runs are summarized in Table 15. The best performances are highlighted in bold.

Table 15. Comparison of MAP of Transliteration Mining and two oracular CLIR systems.

Collection	Oracle-1	Oracle-2	Transliteration Mining	As % of Best Oracle
FIRE 2010 Hindi-English-T	0.3385	0.3374	**0.3390**	100.15
FIRE 2010 Hindi-English-TD	**0.4406**	0.4349	0.4376	99.32
FIRE 2010 Hindi-English-TDN	**0.5026**	0.4942	0.4977	99.03
FIRE 2010 Tamil-English-T	0.3136	**0.314**	0.2815	86.65
FIRE 2010 Tamil-English-TD	**0.3962**	0.3955	0.3621	91.39
FIRE 2010 Tamil-English-TDN	**0.4562**	0.4507	0.4145	90.86

The results presented in Table 15 indicate that the performance Transliteration Mining is nearly equivalent to that of oracular experiments in Hindi-English and fairly close to that for Tamil-English. One curious result needs a bit of explanation: Transliteration Mining outperforms the best oracle in 'Hindi-English-T' setup. On further examination, we found that Transliteration Mining was able to identify two valid equivalents for the transliteratable OOV term '**हिजबुल्लाह**' (a transliteration for the proper noun, '*hezbollah*' or '*hizbolla*') in topic number 77, specifically, '*hizbollahs*' and '*hizbollah*' from the first-pass retrieval; in comparison to the corresponding English topic for the same topic, has only the word '*hezbollah*'. However, we found that a document in the relevant document set contains '*hizbollah*' but not '*hezbollah*'. Thus, Transliteration Mining could outperform the oracle!

6 Conclusion

In this paper, we underlined the need for handling proper and common nouns for improving the retrieval performance of cross-language information retrieval systems. We proposed and outlined two techniques namely Transliteration Mining and Transliteration Generation for handling out of vocabulary (OOV) words to enhance a state of the art baseline CLIR system. Such an enhanced system was used by our team in Microsoft Research India in our participation in the FIRE 2010 shared task [9], for cross-language Hindi-English and Tamil-English retrieval tasks. We presented the performance of our system under various topic configurations, specifically for English monolingual task and two cross-language tasks, Hindi-English and Tamil-English on the standard FIRE 2010 dataset. We showed that each of the two techniques improved retrieval performance, but consistently more so by Transliteration Mining. We also showed specific sample topics to explain and highlight how each technique affects the retrieval performance – positively or negatively, but our experimental evaluation indicate that the overall effect is significantly positive. Finally, we showed that Transliteration Mining performs almost as well as two oracular systems that can identify the transliterations from the equivalent English topic or from the relevant documents from the target collection.

References

1. AbdulJaleel, N., Larkey, L.S.: Statistical transliteration for English-Arabic cross language information retrieval. In: CIKM (2003)
2. Al-Onaizan, Y., Knight, K.: Machine transliteration of names in Arabic text. In: ACL Workshop on Computational Approaches to Semitic Languages (2002)
3. Al-Onaizan, Y., Knight, K.: Translating named entities using monolingual and bilingual resources. In: 40th Annual Meeting of ACL (2002)
4. Ballesteros, L. and Croft, B.: Dictionary Methods for Cross-Lingual Information Retrieval. In: DEXA (1996)
5. Cao, G., Gao, J., Nie, J.Y.: A system to mine large-scale bilingual dictionaries from monolingual Web pages. In: Proceedings of the 11th MT Summit (2007)
6. Chinnakotla, M.K., Vachhani, V., Gupta, S., Raman, K., Bhattacharyya, P.: IITB CFILT @ FIRE 2010: Discriminative Approach to IR. In: Working Notes for the Forum for Information Retrieval Evaluation (FIRE) Workshop (2010)
7. CRF++. http://crfpp.sourceforge.net
8. Demner-Fushman, D., Oard, D.W.: The effect of bilingual term list size on dictionary based cross-language information retrieval. In: 36th Hawaii International Conference on System Sciences (2002)
9. Forum for Information Retrieval Evaluation. http://www.isical.ac.in/~fire
10. Fung, P.: A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In: ACL (1995)
11. Fung, P.: Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In: 3rd Work-shop on Very Large Corpora (1995)
12. He, X.: Using word dependent transition models in HMM based word alignment for statistical machine translation. In: 2nd ACL Workshop on Statistical Machine Translation (2007)
13. Jagarlamudi, J., Kumaran, A.: Cross-Lingual Information Retrieval System for Indian Languages. In: Working Notes for the CLEF 2007 Workshop (2007)
14. Järvelin, A., Järvelin, A.: Comparison of s-gram Proximity Measures in Out-of-Vocabulary Word Translation. In: 15th String Processing and Information Retrieval Symposium (SPIRE) (2008)

15. Joshi, T., Joy, J., Kellner, T., Khurana, U., Kumaran, A., Sengar, V.S.: Crosslingual location search. In: SIGIR (2008) 211-218
16. Khapra, M., Kumaran, A., Bhattacharyya, P.: Everybody loves a rich cousin: An empirical study of transliteration through bridge languages. In: NAACL (2010)
17. Knight, K., Graehl, J.: Machine Transliteration. In: Computational Linguistics (1998)
18. Kraiij, W., Nie, J-Y., Simard, M.: Emebdding Web-based Statistical Translation Models in Cross-Language Information Retrieval. Computational Linguistics (2003)
19. Kumaran, A., Khapra, M., Bhattacharyya, P.: Compositional Machine Transliteration. ACM Transactions on Asian Language Information Processing (TALIP) (2010)
20. Kumaran, A., Khapra, M., Li, H.: Report of NEWS 2010 Transliteration Mining Shared Task. In: 2010 Named Entities Workshop, ACL (2010)
21. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: International Conference on Machine Learning (2001)
22. Li, H., Kumaran, A., Pervouchine, V., Zhang, M.: Report of NEWS 2009 Machine Transliteration Shared Task. In: 2009 Named Entities Workshop: Shared Task on Transliteration (2009)
23. Li, H., Sim, K.C., Kuo, J., Dong, M.: Semantic Transliteration of Personal Names. In: ACL (2007)
24. Majumder, P., Mitra, M., Pal, D., Bandyopadhyay, A., Maiti, S., Mitra, S., Sen, A., Pal, S.: Text collections for FIRE. In: SIGIR (2008)
25. Mandl, T., Womser-Hacker, C.: How do named entities contribute to retrieval effectiveness? In: Cross Language Evaluation Forum Campaign (2004)
26. Mandl, T., Womser-Hacker, C.: The Effect of named entities on effectiveness in crosslanguage information retrieval evaluation. In: ACM Symposium on Applied Computing (2005)
27. Mayfield, J., McNamee, P.: Single n-gram stemming. In: SIGIR (2003)
28. Microsoft Research India. http://research.microsoft.com/en-us/labs/india/
29. Munteanu, D., Marcu, D.: Extracting parallel sub-sentential fragments from non-parallel corpora. In: ACL (2006)
30. Nardi, A., Peters, C.: Working Notes for the CLEF 2007 Workshop (2007)
31. NTCIR. http://research.nii.ac.jp/ntcir
32. Och, F., Ney, H.: A systematic comparison of various statistical alignment models. Computation Linguistics (2002)
33. Peters, C.: Working Notes for the CLEF 2006 Workshop (2006)
34. Pirkola, A., Toivonen, J., Keskustalo, H., Järvelin, K.: Frequency-based identification of correct translation equivalents (FITE) obtained through transformation rules. ACM Transactions on Information Systems (TOIS) 26(1): article 2. (2007)
35. Ponte, J. M., Croft, W.B.: A language modeling approach to information retrieval. In: SIGIR (1998)
36. Porter, M. F.: An algorithm for suffix stripping. In: Program, 14(3):130–137 (1980)
37. Quirk, C., Udupa, R., Menezes, A.: Generative models of noisy translations with applications to parallel fragments extraction. In: 11th MT Summit (2007)
38. Rao, P.R.K., Devi, S.L.: AU-KBC FIRE2010 Submission - Cross Lingual Information Retrieval Track: Tamil- English. In: Working Notes for the Forum for Information Retrieval Evaluation (FIRE) Workshop (2010)
39. Rapp, R.: Automatic identification of word translations from unrelated English and German corpora. In: ACL (1999)

40. Saravanan, K., Udupa, R., Kumaran, A.: Crosslingual Information Retrieval System Enhanced with Transliteration Generation and Mining. In: Working notes for Forum for Information Retrieval Evaluation (FIRE) Workshop (2010)
41. The Cross-Language Evaluation Forum (CLEF). http://clef-campaign.org
42. Udupa, R., Jagarlamudi, J., Saravanan, K.: Microsoft Research India at FIRE2008: Hindi-English Cross-Language Information Retrieval. In: Working notes for Forum for Information Retrieval Evaluation (FIRE) Workshop (2008)
43. Udupa, R., Saravanan, K., Bakalov, A., Bhole, A.: "They Are Out There, If You Know Where to Look": Mining Transliterations of OOV Query Terms for Cross-Language Information Retrieval. In: ECIR (2009)
44. Udupa, R., Saravanan, K., Kumaran, A.: Mining Named Entity Transliteration Equivalents from Comparable Corpora. In: CIKM (2008)
45. Udupa, R., Saravanan, K., Kumaran, A., Jagarlamudi, J.: MINT: A Method for Effective and Scalable Mining of Named Entity Transliterations from Large Comparable Corpora. In: EACL (2009)
46. Virga, P., Khudanpur, S.: Transliteration of proper names in cross-lingual information retrieval. In: ACL Workshop on Multilingual and Mixed Language Named Entity Recognition (2003)
47. Xu, J., Weischedel, R.: Empirical studies on the impact of lexical resources on CLIR performance. Information Processing and Management (2005)
48. Zhai, C., Lafferty, J.: Two Stage Language Models for Information Retrieval. In: SIGIR (2002)
49. Zhai, C., Lafferty, J.: A study of smoothing algorithms for language models applied to information retrieval. In: ACM Transactions on Information Systems, 22(2):179–214 (2004)
50. Zhang, M., Li, H., Kumaran, A., Liu, M.: Report of NEWS 2011 Machine Transliteration Shared Task. In: 2011 Named Entities Workshop, IJCNLP (2011)
51. Zobel, J., Dart, P.: Phonetic string matching: lessons from information retrieval. In: SIGIR (1996)