"*They are out there, if you know where to look*": Mining Transliterations of OOV Query Terms for Cross-Language Information Retrieval

Raghavendra Udupa[1], Saravanan K[1], Anton Bakalov[2], Abhijit Bhole[3]

[1] Microsoft Research India
"Scientia", 196/36, 2nd Main, Sadashivangar, Bangalore 560 080, India.
[2] Harvey Mudd College, 301 Platt Boulevard
Claremont, California 91711-5990
[3] Department of Computer Science and Engineering, IIT Bombay
Powai, Mumbai 400 076, India.

Abstract. It is well known that the use of a good Machine Transliteration system improves the retrieval performance of Cross-Language Information Retrieval (CLIR) systems when the query and document languages have different orthography and phonetic alphabets. However, the effectiveness of a Machine Transliteration system in CLIR is limited by its ability to produce relevant transliterations, i.e. those transliterations which are actually present in the relevant documents. In this work, we propose a new approach to the problem of finding transliterations for out-of-vocabulary query terms. Instead of "generating" the transliterations using a Machine Transliteration system, we "mine" them, using a transliteration similarity model, from the top CLIR results for the query. We treat the query and each of the top results as "comparable" documents and search for transliterations in these comparable document pairs. We demonstrate the effectiveness of our approach using queries in two languages from two different linguistic families to retrieve English documents from two standard CLEF collections. We also compare our results with those of a state-of-the-art Machine Transliteration system.

Keywords: Information Retrieval, Cross-Language Information Retrieval, Out of Vocabulary, Transliteration, Mining, Transliteration Similarity Model

1 Introduction

Cross-Language Information Retrieval (CLIR) systems typically employ translation lexicons for translating the query terms to the language of the document collection. Such translation lexicons are either created by human experts or by automatic processing of parallel corpora. As it is not practically possible to continuously update translation lexicons, they do not guarantee complete coverage of the query terms. For many queries, several of the query terms can not be translated by CLIR systems using

their translation lexicons. Such terms are *out-of-vocabulary* (OOV) from the perspective of CLIR systems. If left untranslated, OOV query terms cause severe degradation in the retrieval performance of the CLIR systems.

Typically, a good number of OOV query terms are proper names and domain specific terms. Proper names and terminology form an open class in any language and new names and terms come to existence and circulation every day. No translation lexicon can ever hope to provide coverage of all names and terms and hence the problem of OOV query terms is a persistent problem for CLIR systems. Such OOV query terms are highly informative and many a time the query itself is centered on them. 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. Further, in countries where English is spoken widely as the second language, code mixing is a natural and commonly observed phenomenon. For instance, in Hindi and Tamil several English common nouns such as **plastic** (**प्लास्टिक, பிளாஸ்டிக்**), **surgery** (**सर्जरी**, **சர்ஜரி**), **banking** (**बैंकिंग, பாங்கிங்**), and **cancer** (**कैंसर**, **கேன்சர்**) have entered the vocabulary and are used in the queries. Such terms are also often not present in translation lexicons.

When the query and document languages share the same orthography and very similar phonetic alphabet, a reasonable strategy to handle OOV terms in the query is to pass them untranslated to the retrieval system. For example, while translating from Spanish to English and vice versa, the name **Richard Nixon** would be translated as it itself. However, when the query language has an orthography and/or phonetic alphabet different from the document language, this simple minded strategy will not work because name translation in this case demands orthographic and phonetic transformation of the terms. Consider, for example, a Hindi-to-English CLIR system where the Hindi language queries are written in the Devanagari script and the English documents in the Latin script. Query terms such as **जेनिका कोस्टेलिक**, **आयर्टन सेना** and **कीमोथेरेपी** must be transliterated to their equivalents in English, namely, **Janica Kostelić**, **Ayrton Senna** and **chemotherapy** which requires orthographic and phonetic transformations.

Machine Transliteration systems are often used to transliterate OOV query terms to the document language. A Machine Transliteration system takes as input a term in the query language and applies phonetic and orthographic transformations on the input string to produce a string in the orthography of the document language. It is well known that the use of a good Machine Transliteration system offers some help in dealing with OOV query terms. Previous works on leveraging Machine Transliteration systems for transliterating OOV query terms have reported statistically significant but not dramatic improvement in the retrieval performance of CLIR systems. The gap between the improvement in the retrieval performance that relevant transliterations provide and that a Machine Transliteration system gives is unfortunately huge.

An important reason why Machine Transliteration systems fail to deliver the maximum is the following: most of the transliterations produced by Machine Transliteration systems are both phonetically and orthographically close to the correct transliterations (as measured by the edit-distance measure) but are not correct. Even

when they are correct, they might not be present in the documents relevant to the query and these documents might contain a slightly different correct variant.

In this paper, we present a novel approach to the problem of transliterating OOV query terms. First, we note that it is not really required to "generate" the transliterations of OOV query terms in order to translate them. The problem can be solved if we can somehow get the transliterations of the OOV query terms. We hypothesize that the best place to look for transliterations of OOV terms of a query are the top results of the CLIR system for the query. We propose a mining algorithm for identifying transliterations in the top CLIR results. The mining algorithm views each query-result pair as a "comparable" document pair. It hypothesizes a match between an OOV query term and a document term in the "comparable" document pair and employs a statistical transliteration similarity model to decide whether the document term is a transliteration of the query term. Transliterations mined in this manner are then used to retranslate the query.

In the remainder of this paper we provide a full exposition of our approach along with results of empirical investigations on queries in two languages from two different linguistic families to retrieve English documents from two standard CLEF collections. We start by discussing some of the important previous research works on transliteration in Section 2. Next we describe our approach in Section 3. We discuss the experimental setup and results of our empirical investigations in Section 4. Finally, we discuss the results and propose some ideas for future investigation in Section 5.

2 Related Work

The problem of translation of OOV query terms has been recognized by several studies to have a significant impact on the performance of CLIR systems [6, 15, 16, 22, 29]. There are two distinct approaches for addressing the problem. The first one focuses on augmenting the translation lexicon by mining comparable corpora [3, 7, 8, 17, 25]. The second approach employs a Machine Transliteration system to transliterate proper nouns [1, 2, 11, 13, 29].

We first discuss some of the approaches for mining a translation lexicon from unrelated and related corpora. Fung hypothesized that words with productive context in one language translate to words with productive context in another language, and words with rigid context translate into words with rigid context. Using this hypothesis, she proposed a measure of the productivity of the context of a word and used it to compile a bilingual translation lexicon from non-parallel English-Chinese corpora [7]. In a related work, Fung used a pattern matching technique to find translations of nouns in general and proper nouns in particular from English-Chinese comparable corpora [8]. Rapp hypothesized that there is a correlation between the patterns of word co-occurrences in corpora of different languages and developed an algorithm for compiling a translation lexicon from non-parallel English-German corpora [25]. Recently, researchers have developed sophisticated algorithms for mining parallel sentences and even parallel sub-sentential fragments from large comparable corpora. Two representative works are that of Munteanu [17] and Quirk

et al [25]. Both works employ a CLIR system for identifying articles with similar content in different languages and proceed to mine parallel fragments from these comparable document pairs. Parallel data mined from comparable corpora is then used to produce a translation lexicon by training statistical word alignment models [20]. Mining Named Entity transliterations from monolingual Web pages and comparable corpora has also been attempted [5, 28]. All of these methods for augmenting translation lexicons require additional data resources such as comparable corpora and only partially solve the problem of OOV query terms.

Machine Transliteration is important not only to CLIR but also to Machine Translation and therefore, it has been studied by researchers from both fields. Knight and Graehl developed finite state transducers for back-transliteration from Japanese to English [11]. Virga and Khudanpur employed statistical machine transliteration techniques to transliterate English names to Chinese and showed statistically significant improvement in the retrieval performance of the HAIRCUT CLIR system [29]. AbdulJaleel and Larkey employed a statistical Machine Transliterator system for English-Arabic CLIR [1]. Joshi et al. proposed a Maximum Entropy based transliteration system and used it for Cross-Language Location Search [10]. Pirkola et al proposed FITE-TRT, a technique for identifying translations of cross-language spelling variants [22]. For a detailed bibliography of research in Machine Transliteration, please see [13].

3 Mining Transliterations from Top CLIR Results

We now describe our approach to address the problem of translating OOV terms. As our approach is designed for terms that could potentially have a transliteration in the document language, we address the OOV problem for query terms that are proper names, domain specific terms, and some common nouns. In the remainder of this paper, we call such terms as transliteratable OOV terms. The problem of finding translations for query terms that are common nouns, adjectives, adverbs, and verbs is beyond the scope of our work.

3.1 Motivating Example

As noted in Section 1, several transliteratable OOV terms, especially names, have low document frequency and are highly informative from the point of view of the query. In many cases, the query itself is centered on such OOV terms. For instance, consider the query from the CLEF 2007 Hindi ad hoc retrieval task that asks for documents on the medals won by the Kostelić siblings in the 2002 Winter Olympics [18]. The query term **कोस्टेलिक** (Kostelić), a proper noun, is unlikely to be present in the Hindi-English translation lexicon and hence will be regarded as OOV by the CLIR system. But from the point of view of expressing the user's information need, this term is indispensable. Any document that discusses either (or both) of the Kostelić siblings can be expected to have some transliteration of the OOV term. In particular, all documents that are relevant to this query will contain some transliteration (e.g. Kostelić or Kostelic) of

the OOV term **कोस्टेलिक**. In general, we may expect to find transliterations of a transliteratable OOV term in many, if not all, of the documents that are relevant to the query. This leads us to the following hypotheses:

Hypothesis 1: The transliterations of most of the transliteratable OOV terms of a query can be found in documents relevant to the query.

Hypothesis 2: If a transliteration of a transliteratable OOV query term is present in some document relevant to the query then it (or a close variation) is present in a majority of the relevant documents.

We tested Hypotheses 1 and 2 on the queries from the CLEF 2006 Hindi-English and CLEF 2007 Hindi-English and Tamil-English ad hoc retrieval tasks [18, 19]. For each transliteratable OOV query term, we searched the corresponding relevant documents for transliterations. The findings of our study are summarized in Table 1 and empirically support the two hypotheses. In all the three cases more than 89% of the transliteratable OOV query terms had at least one transliteration in the relevant documents. Furthermore, whenever a transliteration for an OOV term of a query existed in a relevant document, it (or a close variation) was well expressed in the rest of the relevant documents (Table 1). More than 72% of the transliteratable OOV query terms had a transliteration in at least 50% of the relevant documents.

Table 1: Transliteratable OOV query terms that have transliterations in the relevant documents.

Collection	Transliteratable OOV terms	Terms with transliterations in at least one relevant document	Terms with transliteration in at least 50% of relevant documents
CLEF 2006 (Hindi)	62	58 (94%)	49 (79%)
CLEF 2007 (Hindi)	47	42 (89%)	34 (72%)
CLEF 2007 (Tamil)	43	42 (98%)	39 (89%)

3.2 Towards a Practical Hypothesis

If we knew beforehand the relevant documents for a query, we would not have any need for mining transliterations. After all, the purpose of mining transliterations is to improve the retrieval performance of the CLIR system. On the other hand, Hypothesis 2 says that transliterations are well expressed in the relevant documents. If the CLIR system can bring one or more relevant documents as one of the top results, we may hope to mine transliterations for OOV transliteratable terms. This empirical insight leads us to the following hypothesis which forms the backbone of our approach:

Hypothesis 3: The transliterations of many of the transliteratable OOV terms of a query can be found in the top results of the CLIR system for the query.

3.3 Mining Algorithm

We now develop Hypothesis 3 into a practical method for mining transliterations. Let $\mathbf{q_S}$ be a query and **D** be the top N CLIR results for $\mathbf{q_S}$ with the current translation lexicon **TL**. We pair $\mathbf{q_S}$ with every $\mathbf{d_T} \in \mathbf{D}$ and view the pair ($\mathbf{q_S}$, $\mathbf{d_T}$) as a comparable pair of multilingual documents. We hypothesize a match between each transliteratable OOV term w_S in $\mathbf{q_S}$ and each transliteratable term w_T in $\mathbf{d_T}$. The transliteration similarity of the pair (w_S, w_T) is measured using a transliteration similarity model. We regard all pairs which get a score above a threshold γ as transliteration equivalents. We augment the translation lexicon with the mined transliteration equivalents **TE**. With the resulting translation lexicon $\mathbf{TL}^*$, we get $\mathbf{D}^*$, the top N CLIR results for $\mathbf{q_S}$. We repeat the process for a fixed number of iterations. Tables 2 and 3 give the details of our approach.

Table 2: Algorithm for CLIR with transliterations mining

Algorithm CLIRWithTransliterationsMining

Input:
Number of Iterations M, Translation Lexicon **TL**, Number of Results N, Similarity Threshold γ, Query $\mathbf{q_S}$.

Output:
Top N CLIR results $\mathbf{D}^*$ for $\mathbf{q_S}$ with the augmented Translation Lexicon.

1. $\mathbf{TL}^* = \mathbf{TL}$;
3. For i = 1 to M do
4. $\quad \mathbf{TE_i}$ = MineTransliterationsFromTopResults($\mathbf{TL}^*$, N, γ, $\mathbf{q_S}$);
5. $\quad \mathbf{TL}^* = \mathbf{TL} \cup \mathbf{TE_i}$;
6. End
7. $\mathbf{D}^*$ = TopCLIRResults($\mathbf{q_S}$, N, $\mathbf{TL}^*$);

Table 3: Algorithm for mining transliterations from top results

Algorithm MineTransliterationsFromTopResults

Input:
Translation Lexicon **TL**, Number of Results N, Similarity Threshold γ, Query $\mathbf{q_S}$.

Output:
Transliteration Equivalents **TE** for (some) OOV query terms of $\mathbf{q_S}$.

1. **D** = TopCLIRResults($\mathbf{q_S}$, N, **TL**);
2. **TE** = {};
3. For each OOV term w_S in the query $\mathbf{q_S}$ do

```
4.    If (IsAStopWord(w_S)) then
5.       Continue;
6.    For each document d_T in D do
7.       For each term  w_T in the document d_T do
8.          If (IsAStopWord(w_T)) then
9.             Continue;
10.         If (DoNotHaveComparableLengths(w_S, w_T)) then
11.            Continue;
12.         If (TransliterationSimilarity((w_S, w_T)) > γ ) then
13.            TE = TE  ∪ {(w_S, w_T)};
14.      End
15.   End
16. End
```

3.4 Transliteration Similarity Model

Our transliteration similarity model is an extension of He's W-HMM word alignment model [9] and requires no language-specific knowledge. 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 [20][2]. 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 sparsity 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:

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

Here, t_j (and resp. s_i) denotes the j[th] (and resp. i[th]) character in w_T (and resp. 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}, $j = 1, \cdots, m$. 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 $P(w_T|w_S)$ appropriately transformed.

[2] Although we use a character level hidden alignment model for measuring transliteration similarity, we can, in principle and in practice, employ any reasonable transliteration similarity model including discriminative and/or language-specific models in MineTransliterationsFromTopResults.

4 Empirical Investigations

In this Section, we describe the empirical studies that we conducted to test the central hypothesis of this work.

4.1 Experimental Setup

4.1.1 Data

We conducted our experiments on two English language document collections taken from CLEF: the LA Times 2002 with queries 401-450 (CLEF 2007) and the LA Times 94 + Glasgow Herald 95 with queries 301-350 (CLEF 2006). The topics 401-450 are in Hindi and Tamil and 301-350 are in Hindi [18, 19]. An English version of the queries is also available. As the collections and topics are from past years, their relevance judgments are also available. We used all the three fields (title, description, and narration) of the CLEF topics.

4.1.2 Dictionaries

We used statistical dictionaries for both Hindi-English and Tamil-English CLIR. We generated the dictionaries by training statistical word alignment models on Hindi-English parallel corpora (~55K parallel sentences) and Tamil-English parallel corpora (~40 K parallel sentences) using the GIZA++ tool [20]. We used 5 iterations of IBM Model 1 and 5 iterations of HMM [20]. We retained only the top 4 translations for every source word.

4.1.3 CLIR System

We used a query likelihood based ranking approach for ranking the documents [23, 30]. We used only the textual content of the documents for indexing and indexed only non-empty documents. We removed stop words from the text while indexing and stemmed the words using the Porter stemmer [24].

4.1.4 Transliteration Similarity Model

We trained Hindi-English and Tamil-English transliteration similarity models on 16 K parallel single word names in Hindi-English and Tamil-English respectively. We did 15 iterations of EM.

4.1.5 Parameters

We used the following setting in our experiments:
Number of Iterations: M = 2.
Number of Top Results: N = 150 for the first iteration, 50 for the second iteration.
Transliteration Similarity Threshold: γ = 1.5 for CLEF 2007 collection and 1.0 for CLEF 2006 collection.

4.2 Mining Results

We did two iterations of mining for each collection and the results are presented in Table 4. As can be noted from Table 4, the mining results provide strong evidence for Hypothesis 3. For Hindi-English direction, we could successfully find at least one transliteration for more than 61% of the transliteratable OOV terms present in the queries. For Tamil-English direction, the percentage of transliteratable OOV terms for which we mined a transliteration was a respectable figure of 37. Furthermore, on an average, we mined more than 1 transliteration for each transliteratable OOV term. This brings in some amount of query expansion automatically. We also observed that two iterations of mining provided best results on two collections. We did not go for more number of iterations because the returns diminished after two iterations.

Table 4: Transliterations mined from top CLIR results.

Collection	Transliteratable OOV terms	Iteration 1		Iteration 2	
		Terms for which at least one valid transliteration was mined	Valid transliterations mined	Terms for which at least one valid transliteration was mined	Valid transliterations mined
CLEF 2006 (Hindi)	62	35 (56%)	50 (1.43/term)	38 (61%)	55 (1.45/term)
CLEF 2007 (Hindi)	47	30 (64%)	42 (1.40/term)	30 (64%)	45 (1.50/term)
CLEF 2007 (Tamil)	43	14 (33%)	23 (1.64/term)	16 (37%)	25 (1.56/term)

4.3 CLIR Results

The query likelihood CLIR system is the baseline for all CLIR experiments. In order to compare the performance of CLIRWithTransliterationsMining with a state-of-the-art Machine Transliteration based CLIR system, we used the MaxEnt transliterator described in [10] for transliterating OOV query terms. We used only the top 4 transliterations. The results of the CLIR runs are summarized in Table 5. We observed improvements in the retrieval performance with both CLIRWithTransliterationsMining and MaxEnt Transliterator. However, CLIRWithTransliterationsMining performed the best.

We also built two oracular CLIR systems to determine a reasonable upper bound for the retrieval performance of CLIRWithTransliterationsMining. The first oracular system made use of gold transliterations from the English queries. The second oracular system made use of gold transliterations from the relevant documents. The oracular CLIR systems used the same dictionaries as the CLIRWithTransliterationsMining system. The results are summarized in Table 6.

Table 5: Comparison of the retrieval performances of the baseline, MaxEnt Transliterator, and CLIRWithTransliterationsMining systems. The evaluation measure is Mean Average Precision (MAP). Stars indicate statistically significant differences with 95% confidence according to paired t-test.

Collection	Baseline	MaxEnt Transliterator	% change over baseline	CLIRWithTransliterationsMining			
				Iteration 1	% change over baseline	Iteration 2	% change over baseline
CLEF 2006 (Hindi)	0.1463	0.157	+7.31*	0.2476	+69.24*	**0.2527**	**+72.73***
CLEF 2007 (Hindi)	0.2521	0.2761	+9.52	**0.3389**	**+34.43***	0.3380	+34.07*
CLEF 2007 (Tamil)	0.1848	0.2024	+9.52	0.2270	+22.84*	**0.2279**	**+23.32***

Table 6: Comparison of the retrieval performance of the best CLIRWithTransliterationsMining system with two oracular CLIR systems. The evaluation measure is Mean Average Precision (MAP).

Collection	Oracle-1	Oracle-2	Best Mining	As % of Best Oracle
CLEF 2006 (Hindi)	0.3022	0.3076	0.2527	82
CLEF 2007 (Hindi)	0.3696	0.3770	0.3389	90
CLEF 2007 (Tamil)	0.2761	0.2854	0.2279	80

We noticed that both oracular systems gave better retrieval performance than CLIRWithTransliterationsMining but our system achieved more than 80% of the best oracular system. Next, we removed incorrect transliterations from the output of MineTransliterationsFromTopResults and evaluated the retrieval performance of CLIRWithTransliterationsMining. We observed a small, but not statistically significant, improvement in the retrieval performance. This means that incorrect transliterations mined by our algorithm do not significantly hurt the retrieval performance.

4.3.1 Performance Analysis

Figure 1 shows the query-level difference in the Average Precision between the baseline and our method on the three test collections. We see that in each test several topics have profited from the mined transliterations.

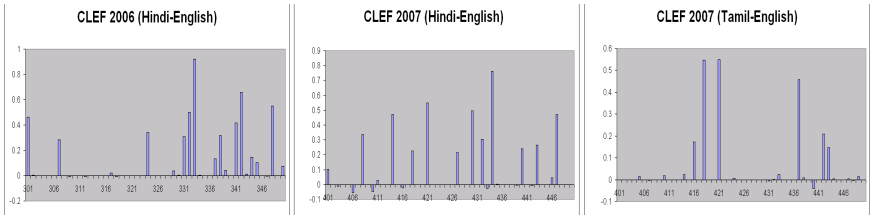

Figure1. Differences in Average Precision between the baseline and CLIRWithTransliterationsMining.

5 Conclusions and Future Work

We proposed a novel approach for the problem of OOV query terms in CLIR based on the key hypothesis that the top CLIR results for a query have the transliterations of many of the OOV terms in the query. We developed this hypothesis into a practical method. We provided experimental evidence for our hypothesis and showed that it results in highly impressive gains in the retrieval performance. We provided good empirical upper bounds for the retrieval performance of our system and showed that our performance is quite close to these upper bounds. We also compared our performance with that of a state-of-the-art transliterator.

One promising direction of future work is the use of a good stemmer for inflectional languages such as Tamil. A good stemmer is likely to improve the performance of our algorithm for Tamil-English CLIR. For instance, topic 403 in CLEF 2007 contains the term **போலீஸ்காரர்களாக** (like the Police) which is an inflected form of **போலீஸ்** (police) which is present in several of the relevant documents for the topic. Another interesting possibility is the use of a discriminative classifier in the mining algorithm. Finally, it would be interesting to use a Machine Transliteration system along with our system.

6 Acknowledgments

We thank Jagadeesh Jagarlamudi, Joseph Joy, A.Kumaran, Sandipan Dandapat, Doug Oard, Jian-Yun Nie, Sanjeev Khudanpur, and Paul McNamie for helpful comments.

References

1. AbdulJaleel, N. and Larkey, L.S.: Statistical transliteration for English-Arabic cross language information retrieval. In Proceedings of CIKM 2003.
2. Al-Onaizan, Y. and Knight, K.: Machine transliteration of names in Arabic text. In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages (2002).

3. Al-Onaizan, Y. and Knight, K.: Translating named entities using monolingual and bilingual resources. In Proceedings of the 40th Annual Meeting of ACL (2002).
4. Ballesteros, L. and Croft, B. 1998. Dictionary Methods for Cross-Lingual Information Retrieval. In Proceedings of DEXA'96.
5. Cao, G., Gao, J., and Nie, J-Y.: A system to mine large-scale bilingual dictionaries from monolingual Web pages. In Proceedings of the 11th MT Summit (2007).
6. Demner-Fushman, D. and Oard, D. W.: The effect of bilingual term list size on dictionary-based cross-language information retrieval. In Proceedings of the 36th Hawaii International Conference on System Sciences (2002).
7. Fung, P.: A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In Proceedings of ACL 1995.
8. Fung, P.: Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In Proceedings of the 3rd Work-shop on Very Large Corpora (1995).
9. He. X.: Using word dependent transition models in HMM based word alignment for statistical machine translation. In Proceedings of 2nd ACL Workshop on Statistical Machine Translation (2007).
10. Joshi et al.: Cross-Lingual Location Search. In Proceedings of SIGIR 2008.
11. Knight, K. and Graehl, J.: Machine Transliteration. Computational Linguistics (1998).
12. Koehn, P. and Knight, K. Learning a translation lexicon from monolingual corpora. In Proceedings of Unsupervised Lexical Acquisition (2002).
13. Li et al.: Semantic Transliteration of Person Names. In Proceedings of ACL 2007.
14. McNamee, P. and Mayfield, J..: Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources. In Proceedings of SIGIR 2001.
15. Mandl, T. and Womser-Hacker, C.: How do named entities contribute to retrieval effectiveness? In Proceedings of the 2004 Cross Language Evaluation Forum Campaign.
16. Mandl, T. and Womser-Hacker, C.: The Effect of named entities on effectiveness in cross-language information retriev-al evaluation. In ACM Symposium on Applied Computing 2005.
17. Munteanu, D. and Marcu D.: Extracting parallel sub-sentential fragments from non-parallel corpora. In Proceedings of the ACL 2006.
18. Nardi, A. and Peters, C. (ed.): Working Notes for the CLEF 2007 Workshop.
19. Nardi, A. and Peters, C. (ed.): Working Notes for the CLEF 2006 Workshop.
20. Och, F. and Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics, 2003.
21. Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K. and Jarvelin, K.: Fuzzy translation of cross-lingual spelling variants. In Proceedings of SIGIR 2003.
22. Pirkola, A., Toivonen, J., Keskustalo, H., and Jarvelin, K.: FITE-TRT: A high quality translation technique for OOV words. In Proceedings of SAC'06.
23. Ponte, J.M. and Croft, W. B.: A language modeling approach to information retrieval. In Proceedings of SIGIR 1998.
24. Porter, M. F.: An algorithm for suffix stripping. Program 14(3) pp 130-137, 1980.
25. Quirk, C., Udupa, R. and Menezes, A.: Generative models of noisy translations with applications to parallel fragments extraction. In Proceedings of the 11th MT Summit (2007).
26. Rapp, R.: Automatic identification of word translations from unrelated English and German corpora. In Proceedings of ACL'99.
27. Xu, J. and Weischedel, R.: Empirical studies on the impact of lexical resources on CLIR performance. Information Processing and Management (2005).
28. Udupa, R., Saravanan, K. and Kumaran, A.: Mining Named Entity Transliteration Equivalents from Comparable Corpora. In Proceedings of CIKM 2008.
29. Virga, P. and Khudanpur, S.:Transliteration of proper names in cross-lingual information retrieval. In Proceedings of the ACL Workshop on Multilingual and Mixed Language Named Entity Recognition (2003).
30. Zhai, C. and Laffery, J.: A study of smoothing algorithms for language models applied to information retrieval. ACM Trans. On Inf. Sys. 22(2) pp 179-214 (2004).
31. Zhou, D., Turan, M. and Brailsford, T.: Ambiguity and unknown term translation is CLIR. In Working Notes for the CLEF 2007 Workshop.