# Transliteration Retrieval Model for Cross Lingual Information Retrieval

Ea-Ee Jan[1], Shih-Hsiang Lin[1,2], and Berlin Chen[2]

[1] IBM T.J. Watson Research Center, NY 10598, USA
{ejan,shihhlin}@us.ibm.com
[2] Computer Science and Information Engineering, National Taiwan Normal University, Taipei, Taiwan
{shlin,berlin}@csie.ntnu.edu.tw

**Abstract.** The performance of transliteration from a source language to a target language builds the ground work in support of proper name Cross Lingual Information Retrieval (CLIR). Traditionally, this task is accomplished by two separate modules: transliteration and retrieval. Queries are first transliterated to target language using one or multiple hypotheses. The retrieval is then carried out based on translated queries. The transliteration often results in 30-50% errors with top 1 hypothesis, thus leading to significant performance degradation in CLIR. Therefore, we proposed a unified transliteration retrieval model that incorporates the transliteration similarity measurement into the relevance scoring function. In addition, we presented an efficient and robust method in similarity measurement for a given proper name pair using the Hidden Markov Model (HMM) based alignment and a Statistical Machine Translation (SMT) framework. Experimental data showed significant results with the proposed integrated method on the NTCIR7 IR4QA task, which demonstrated a greater flexibility and acceptance in transliteration.

**Keywords:** cross lingual information retrieval (CLIR), transliteration, retrieval model, statistical machine translation (SMT), NTCIR.

## 1 Introduction

Proper name transliteration, the pronunciation-based translation of a proper name from a source language to a desired target language, is important to many multilingual natural language processing tasks, including Cross Lingual Information Retrieval (CLIR), multilingual spoken document retrieval and Machine Translation (MT). Traditionally, transliteration is achieved by human translators using hand-crafted translation lexicon and translation rules. However, due to the fast growth of multilingual information, more and more proper names become undefined in the lexicon. These undefined proper names can be referred as Out-Of-Vocabulary (OOV) word, and potentially introduce harm to the system performance, especially in CLIR task. Since proper names usually carry distinctive information, when not carefully handled, the mean average precision (mAP) can reach 50% degradation [1].

P.-J. Cheng et al. (Eds.): AIRS 2010, LNCS 6458, pp. 183–192, 2010.

Transliteration between languages that use irrelevant alphabets and phonemes, e.g. English and Chinese or English and Arabic is especially challenging [1-4]. Due to the pronunciation differences between the source and target languages, particular phonemes in the source language can either have no corresponding phoneme or have a few close phonemes in the target language. These pronunciation differences usually yield “many to one”, “one to many” or “one to none” phoneme mappings. Thus, one proper name can be transliterated into multiple foreign names with similar pronunciations. Differential alignment in source language can also lead to different transliteration outcomes in the target language. In addition, because of homophones, a single proper name can be transliterated to multiple foreign names with identical pronunciation. A typical example of multiple transliteration is that “Osama Bin Laden” can be transliterated into “奧薩瑪 賓拉登”, “奧薩瑪 本拉丹”, “奧薩瑪 本拉登” and so on. These varieties can significantly degrade the CLIR performance when a different set of transliterations was used by the query and its relevant documents. Despite that the state-of-the-art statistical machine translation (SMT) techniques have been developed, the proper name transliteration remains as a challenging task due to its diversity. Hence, alternative methodology that targets multiple transliterations is indispensible.

Proper name CLIR was mostly tackled by two essential steps [1, 4-6]: (1) one or multiple alternative transliterations were generated for each unknown word using transliteration techniques; (2) the expanded alternative transliterations were used for the retrieval module. However, two potential problems remained unaddressed. Firstly, transliteration does not always produce the exact spelling variants used in the document collection. An approximate string matching technique is usually required to alleviate this drawback. Secondly, the retrieval performance will be compromised if inaccurate or confusing transliterations occur in multiple alternative transliterations. To our best knowledge, no unified framework has been used to seamlessly integrate transliteration and retrieval model. Therefore, we proposed a novel transliteration retrieval model that incorporates a transliteration similarity function into the relevance scoring mechanism as a single component. The proposed framework computes the document relevant score of a given query without translation of any words, even the quires and documents are represented in two irrelevant languages. In this framework, we employed an efficient and robust method to measure the similarity of a given proper name pair using a Hidden Markov Model (HMM) based alignment.

## 2 Background and Related Work

Related work in CLIR can be roughly grouped into two major categories [7]: (1) translating the query into the language used in the document collection and (2) incorporating the statistical translation model into the retrieval model. For the first category, after the queries have been translated into the language used in the document collection, the remaining retrieval procedure can be derived from the standard ad hoc retrieval [3-6]. However, this approach relies on a high quality machine translation method. In the second category, researches have attempted to apply the statistical Language Model (LM) to CLIR [8-10]. The LM aims to capture the regularity in human natural language and to quantify the acceptability of a given word sequence.

The basic idea is that a document is deemed to be relevant to a query if its corresponding document language model is more likely to generate the query. A probabilistic language model is developed explicitly for each individual document in the collection. Therefore, the translation probability $P(t \mid s)$ can be naturally integrated into the retrieval model where $s$ is the source term and $t$ is the target term. However, the diversity of the proper name transliteration makes it difficult to have a translation model that handles all the possible proper name translation.

The proper name transliteration was often modeled by the SMT framework. Virga et al. applied statistical machine translation models to translate English names into Chinese characters for Mandarin spoken document retrieval [4]. Knight and Graehl proposed a generative transliteration model for Japanese and English using finite state transducers [11]. Meng et al. developed an English-Chinese Named Entity (NE) transliteration technique using a pronunciation lexicon and phonetic mapping rules [3]. Most of the above statistical machine translation approaches were based on the IBM noisy source-channel model framework [12]. Instead of using the noisy source-channel model, Gao et al. proposed a direct modeling approach to estimate the posterior probability using phoneme chunks as the contextual features [13]. Li et al. proposed direct orthographic mapping with a joint source-channel model for proper name transliteration [14]. Kumaran and Kellner [15] also implemented a generic transliteration framework using an approach similar to [14]. On the other hand, due to the challenge of machine transliteration, additional data source such as comparable corpora and web were explored to improve the performance [16-17]. All of these approaches had a scoring mechanism to test how likely a given pair of names in source and target languages is the transliteration of each other. This is a key component and is the aspect we focus on in this paper.

One popular approach for proper name CLIR is to translate queries from the source language to the target language. IR module is then applied for document retrieval. By this approach, the CLIR is addressed by two separated tasks. For the LM based IR, a probabilistic language model is developed explicitly for each individual document in the collection. The translation probability $P(t|s)$ is needed for retrieval model. However, proper name transliteration is a challenge task. The $P(t|s)$ can be less accurate. In this paper, we propose a unified framework which integrates the proper name transliteration similarity measurement into the retrieval model.

## 3 Our Approach

In this section, we first describe the baseline retrieval model and its extension to the proposed transliteration retrieval model. We then present our approach to calculate proper name transliteration similarity in latter section.

### 3.1 Retrieval Model

In language model based IR, a probabilistic generative framework is used for ranking each document $D$ in the collection by a given query $Q$. This concept can be described by $P(D \mid Q)$ [18]. The ranking criterion is usually approximated by the likelihood of $Q$ generated by $D$, i.e., $P(Q \mid D)$. By this approach, each document is treated as a probabilistic language model for generating the query. If $Q$ is treated as a sequence of words, $Q = w_1 w_2 \cdots w_N$, which is assumed to be independent of each other given $D$, and the word order is assumed to be irrelevant, (the "*bag-of-words*" assumption), the relevance measure $P(Q \mid D)$ can be decomposed as a product of the probabilities of the query words generated by the document:

$$\begin{aligned} P(Q \mid D) &= \prod_{w \in Q} P(w \mid D)^{c(w,Q)} \\ &\propto \sum_{w \in Q} c(w,Q) \log P(w \mid D) \end{aligned} , \tag{1}$$

where $c(w,Q)$ is the number of times that each distinct word $w_i$ occurs in $Q$. The document ranking is now simplified by the document model $P(w \mid D)$. The simplest way to construct $P(w \mid D)$ is using the unigram LM, where each document in the collection can respectively offer a unigram distribution for observing a query word.

A similar notion can be applied to CLIR by introducing the translation probability $P(w_e \mid w_f)$ into the formulation where $w_e$ is an English query term and $w_f$ is a Chinese document term[1] [19]. The retrieval model for CLIR given the English query $Q_e$ and the Chinese document $D_f$ can be formulated as:

$$\begin{aligned} P(Q_e \mid D_f) &= \prod_{w_e \in Q_e} P(w_e \mid D_f)^{c(w_e,Q_e)} \\ &\propto \sum_{w_e \in Q_e} c(w_e,Q_e) \log\left( \sum_{w_f \in D_f} P(w_e \mid w_f) P(w_f \mid D_f) \right) \end{aligned} . \tag{2}$$

Due to the OOV and diversity natural of the transliteration, it can be a challenge to estimate the translation probability of $P(w_e \mid w_f)$. We therefore redefine the scoring function as

$$P(Q_e \mid D_f) \propto \sum_{w_e \in Q_e} c(w_e,Q_e) \log\left( \sum_{w_f \in D_f} Transliteration(w_e, w_f) P(w_f \mid D_f) \right) , \tag{3}$$

where $Transliteration(w_e, w_f)$ is an approximated transliteration probability obtained by the following sigmoid operation:

$$Transliteration(w_e, w_f) = \frac{1}{1 + \exp(-\gamma \cdot sim(w_e, w_f) + \beta)} , \tag{4}$$

The weight $\gamma$ and $\beta$ describe the steepness and central of the sigmoid function, respectively. The $Sim(w_e, w_f)$ denotes the similarity between the query term $w_e$ and the document term $w_f$. This sigmoid function is introduced to converts the similarity measure, $Sim(w_e, w_f)$, into a probability measure. The idea underlying the proposed model is that the greater the similarity score implies the greater likelihood of an appropriate transliteration.

---

[1] In the SMT community, *e* and *f* are represented as English and foreign language, respectively. In this paper, *e* denotes English and *f* is Chinese.

### 3.2 Transliteration Similarity $Sim(w_e, w_f)$

For a given proper name pair, one from the source language and the other from the target language, our goal here is to explore approaches with reliable similarity scoring mechanism which yields high accuracy with low computational complexity. For ease of illustration, we choose English and Chinese name transliteration for this study.

Intuitively, proper name transliteration "translates" a proper name from the source language to the target language based on pronunciation. Phonetic based edit distance measures should provide a good evaluation method. However, the source and target language can have very different base phone sets. One has to convert these phone sets to a unified phone set for edit distance calculation. Due to pronunciation differences between the source and target languages, the phone set mapping can be "one to many", "many to one" or "one to none". Additionally, pronunciation difference between some phone pairs can be more significant than the other pairs. For example, the transliteration difference caused by /t/ and /d/ should incur less error than the differences cause by /t/ and /e/. Instead of treating all errors with unique cost, we need a similarity measure rather than phonetic based edit distance. Some linguistic background may help hand craft the phone set mapping rules with weights. Instead, we use a phonetic based SMT framework to derive the mapping rules.

We propose a left-to-right discrete HMM based alignment to measure the similarity for a given proper name pair. Prior to alignment, the proper name pairs are converted to phone sequences. In the discrete HMM alignment, we treat each phone to be aligned as a state which is characterized by a multinomial distribution. The emission probability for each state is the conditional probability of phone in target language given by phone in source language $p(f \mid e)$, or by a null probability model. The null probability model is presented by either $p(f \mid \phi)$ or $p(\phi \mid e)$. The phrase-based SMT framework can be utilized to derive the state emission probability $p(f \mid e)$. The valid state transition is from left to right with self looping, and with the maximum jump of two states. A uniform probability is used for all valid state transition. The best path can be calculated by the dynamic time wrapping (DTW). The similarity score is the best alignment score normalized by the total length.

The parallel training corpus for SMT to derive $p(f|e)$ is organized as such that the English phone sequences (in English phone set) are paired with their corresponding Chinese phone sequences (in Chinese phone set). Words with multiple pronunciations are fully enumerated. The SMT models are developed by a commonly used recipe [22]. The phrase table $p(f \mid e)$ is used as the emission probability for the HMM alignment. The phrase table size is set to 2. Thus, one or two phone sequences in the source language can mapped to one or two phone sequences in the target language and vice versa. Please refer to [24] for more details.

## 4 Experimental Setup

Two consecutive experiments were setup to verify the proposed framework. The first experiment is to assess the performance of proper name transliteration similarity measurement. To this end, a parallel corpus consisting of English and Chinese proper name pairs was extracted from the people section of the multilingual Wikipedia. Approximately 3,000 pairs were used for training and 300 pairs for testing. To evaluate the robustness of the proposed similarity measurement, match and unmatched conditions were both tested. The 300 pairs were used as a matched condition test. A separated 1,000 unmatched test pairs were created randomly from the 300 matched pairs. The English and Chinese pronunciations were obtained by the IBM voice toolkits. Multiple pronunciations for a given word were considered to be uniformly distributed. All possible combinations of pronunciation were created in both training and test set.

The second experiment is document retrieval using the data compiled from the NTCIR-7 Information Retrieval for Question Answering (IR4QA) task [20]. IR4QA evaluates the performance of document retrieval using Average Precision (AP) metrics. This task is embedded in the context of cross language question answering. In this study, we used the EN-CS (English to Simplified Chinese) subtask, which includes 545,162 documents and 97 queries, for the retrieval experiments. Of the original 97 queries, 10 proper name related queries (listed in Table 2) were extracted to verify our proposed method. The retrieval results were presented by AP.

## 5 Results and Discussion

We first evaluated the performance of the proper name transliteration similarity measurement. The phonetic based transliteration SMT system was developed. The log probability of e2f and f2e from phrase tables were used to calculate the similarity score for both matched and unmatched proper name pair test set. A larger value represents a higher similarity. If the similarity score for a matched pair is lower than a given threshold, this pair is falsely rejected. If the similarity score for an unmatched pair is higher than the threshold, this pair is falsely accepted. The performance metrics can be evaluated by the Equal Error Rate (EER), where the error rate of the false rejection of matched pairs and the false acceptance of unmatched pairs are identical. The EER for our proposed method was 3.47%. This low EER provides a good foundation for proper name CLIR. For comparison, the transliteration employing the commonly used SMT framework was also explored. The translation of a given phone sequence from the source language to the target language was performed. The similarity of the translated phone sequence and true target phone sequence was calculated using BLEU score [23]. While setting the phrase table size equal to 8, the SMT framework yielded the best performance at the EER of 7.1%, which is substantially higher than that of our approach. We also compared the edit distance between spelling of English name and Pinyin of Chinese names. The EER for the orthographic based edit distance is 22%. Table 1 summarizes these results.

**Table 1.** Equal error rates (EERs) for various similarity measures

| Approach | EER(%) |
|---|---|
| Orthographic Edit distance | 22.27% |
| Phonetic Based SMT | 7.10% |
| Proposed Method | 3.43% |

**Table 2.** Baseline retrieval results (in AP) for NTCIR-7 ACLIA EN-CS

| Topic | Query | Baseline | NTCIR7(Avg.) |
|---|---|---|---|
| ACLIA-CS-T42 | 本拉登 (Bin Laden) | 0.0585 | 0.1628 |
| ACLIA-CS-T43 | 罗讷尔多 (Ronaldo) | 0.8664 | 0.6319 |
| ACLIA-CS-T55 | 哈塔米 (Khatami) | 0.7843 | 0.5054 |
| ACLIA-CS-T61 | 克林顿 (Clinton)<br>莱温斯基 (Lewinsky) | 0.6659 | 0.4514 |
| ACLIA-CS-T338 | 苏哈托 (Suharto) | 0.7747 | 0.5837 |
| ACLIA-CS-T339 | 普里马可夫 (Primakov) | 0.8304 | 0.6226 |
| ACLIA-CS-T340 | 农德孟 (Nong Duc Manh) | 0.9485 | 0.6935 |
| ACLIA-CS-T367 | 拉纳利 (Ranariddh) | 0.7103 | 0.4774 |
| ACLIA-CS-T376 | 奥尼尔 (O'Neal) | 0.6275 | 0.3967 |
| ACLIA-CS-T379 | 郑肯 (Duncan) | 0.6442 | 0.3685 |
| Avg. | - | 0.6911 | 0.4894 |

For document retrieval, we first evaluated the retrieval model (cf. Eq. (1)) on the monolingual task (CS-CS) using entire 97 queries. The unigram document language model $P(w \mid D)$ is constructed with Dirichlet smoothing where the smoothing parameter $\mu$ is determined by maximizing the leave-one-out log likelihood of the entire document collection [21]. The AP of this approach is 0.5764. The best performance for this query set from the NTCIR report is 0.6184 and the average AP from all participated systems is 0.4276. Our monolingual baseline retrieval model is comparable to the best systems of the NCTIR-7 report.

We then performed retrieval experiment using the 10 proper name query topics. Our baseline yielded overall AP of 0.6911, which outperformed the average AP[2] of 0.4894 from all systems (cf. Table 2). However, while examining the AP of each individual query topic, our baseline performed is worse than the average for the query ACLIA-CS-T42. The reason is that the proper name "Bin Laden" has been transliterated as "本拉登", "本拉丹", and so on, in the document collection. This multiple transliterations degrade the performance of our baseline, which uses words as the indexing unit. To address the problem of ACLIA-CS-T42 query topic, it is necessary to extend the baseline to our proposed approach, which integrates the transliteration similarity function into the retrieval model (cf. Eq. (3)). To evaluate the performance of proposed approach, we needed the document collection equipped with controlled multiple transliterations. We created a homogenous name list for those proper names used in the test query topics and uniformly replaced those names into the original document collections. The homogenous names list and associated Pinyin are shown in Table 3. The baseline performance for each original query against the new synthetic document collection was presented at Table 3 (denoted as "**Synthetic**"). When a proper name has $n$ alternatives, the performance degradation will be approximately $(n/n+1)$.

[2] The best performance for individual query is not available from NTCIR reports. We can only compare our baseline with the average AP.

**Table 3.** The homogenous name list and retrieval results (in AP) on the synthetic collection

| **Original** | **Alternatives** | **Synthetic** |
|---|---|---|
| 罗讷尔多 (Ronaldo)<br>/luo na er duo/ | 罗讷度 罗讷多 罗讷尔度 朗拿度<br>/luo ne du/ /luo ne duo/ /luo ne er du/ /lang na du/ | 0.1284 |
| 哈塔米 (Khatami)<br>/ha ta mi/ | 卡塔米 汉塔米 哈他米 肯塔米<br>/ka ta mi/ /han ta mi/ /ha ta mi/ /ken ta mi/ | 0.1261 |
| 克林顿 (Clinton)<br>/ke lin dun/ | 柯林顿 科林顿 可林顿<br>/ke lin dun/ /ke lin dun/ /ke lin dun/ | 0.2126 |
| 莱温斯基 (Lewinsky)<br>/lai wen si ji/ | 路文斯基 李文斯基 莱文斯基<br>/lu wen si ji/ /li wen si ji/ /lai wen si ji/ | |
| 苏哈托 (Suharto)<br>/su ha tuo/ | 撒哈托 苏哈尔托 撒哈尔托 史哈托<br>/sa ha tuo/ /su ha er tuo/ /sa ha er tuo/ /shi ha tuo/ | 0.1220 |
| 普里马可夫 (Primakov)<br>/pu li ma ke fu/ | 普林马可夫 普利马可夫 普利马科夫 普利马克夫<br>/pu li ma ke fu/ /pu lin ma ke fu/ /pu li ma ke fu/ /pu li ma ke fu/ | 0.1418 |
| 农德孟(Nong Duc Manh)<br>/nong de meng/ | 侬德猛 农德猛 侬得猛<br>/nong de meng/ /nong de meng/ /nong de meng/ | 0.2124 |
| 拉纳利 (Ranariddh)<br>/la na li/ | 拉那烈 拉那利 拉纳瑞德 拉那瑞<br>/la na lie/ /la na li/ /la na rui de/ /la na rui/ | 0.4564 |
| 奥尼尔 (O'Neal)<br>/ao ni er/ | 欧尼尔 澳尼尔 欧尼而 欧尼耳<br>/ou ni er/ /ao ni er/ /ou ni er/ /ou ni er/ | 0.1625 |
| 邓肯 (Duncan)<br>/deng ken/ | 当肯 丹肯 郑肯<br>/zheng ken/ /dang ken/ /dan ken/ | 0.1556 |

We then evaluated the proposed method against the synthetic collection document. We first assumed a perfect similarity function (Eq. (4)): $Sim(W_e, W_f)=1$ if a pair of words is in the homogenous name list; otherwise, $Sim(W_e, W_f)=0$. This ideal function can be the Oracle results of our approach (denoted by "**Oracle**" in Table 4). The Oracle results outperform the baseline results for query topics "ACLIA-CS-T42" and "ACLIA-CS-T367", which suggests that the original document collection may have multiple transliterations for these two query topics. This approach also slightly outperforms the baseline in a few query topics. This performance improvement can stem from the segmentation error. For example, if "科林顿路"is included in the document lexicon, it might be a mis-segmentation of "科林顿", whereas our approach is able to retrieve this document. However, this experiment is impractical since the complete list of transliterations for the document collection is unknown.

Next, we manually tagged 2,000 candidate proper names from the document collection for similarity function evaluation. These 2,000 candidate proper names are related to the homogenous name list. At least one character from the candidate proper name has a similar pronunciation to the homogenous name list. This test data was used to evaluate the robustness of our approach against the false acceptance of OOVs. The similarity scores for these proper names were calculated. The weights of $\gamma$ and $\beta$ used in sigmoid operation (cf. Eq. (4)) were optimized based on the experiments shown in Table 1 (with an EER of 4.3%). The corresponding results were shown in Table 4 (denoted by "**Taggers**"). The average of all APs was degraded slightly from 0.7254 to 0.6516, while compared to the Oracle results.

Last, we evaluated the worse scenario by using the vocabulary from the complete set of lexicon with length of 2 to 5, to calculate similarity function. It resembles a scenario when the Named Entities (NE) tagger is not available. This test scenario created the maximum number of false acceptance of OOVs. The average AP for this experiment drops from the Oracle performance of 0.7254 to 0.4040 (denoted by "**Worst**"), but it is still much better than that of 0.1776 obtained by the baseline retrieval approach on synthetic document collection. The results clearly showed that our proposed transliteration retrieval model and the low EER of similarity function can properly handle the multiple proper name transliteration problems. Even without the NE tagger (cf. "**Worst**" in Table 4), the proposed method provides significant improvements over the baseline of the synthetic document collection.

**Table 4.** Retrieval results (in AP) for the simulated document collection

| Topic | Synthetic | Oracle | Tagger | Worst |
|---|---|---|---|---|
| ACLIA-CS-T42* | 0.0585 | 0.1391 | 0.0777 | 0.0522 |
| ACLIA-CS-T43 | 0.1284 | 0.8663 | 0.3870 | 0.1497 |
| ACLIA-CS-T55 | 0.1261 | 0.8014 | 0.7936 | 0.2112 |
| ACLIA-CS-T61 | 0.2126 | 0.6899 | 0.6884 | 0.6977 |
| ACLIA-CS-T338 | 0.1220 | 0.7749 | 0.6642 | 0.0415 |
| ACLIA-CS-T339 | 0.1418 | 0.8304 | 0.8307 | 0.8360 |
| ACLIA-CS-T340 | 0.2124 | 0.9485 | 0.9176 | 0.9117 |
| ACLIA-CS-T367* | 0.4564 | 0.9319 | 0.8931 | 0.8784 |
| ACLIA-CS-T376 | 0.1625 | 0.6276 | 0.6293 | 0.2613 |
| ACLIA-CS-T379 | 0.1556 | 0.6442 | 0.6340 | 0.0002 |
| Avg. | 0.1776 | 0.7254 | 0.6516 | 0.4040 |

## 6 Conclusion

Due to the diversity of proper name transliteration, the transliteration accuracy can be impaired, leading to performance degradation in proper name CLIR. We proposed a unified transliteration retrieval framework which integrates the transliteration similarity measurement into the relevance scoring function. Instead of performing proper name transliteration, a transliteration similarity function is used in our framework. The EER of the proposed similarity function can be as low as 3.5%, which reduces the negative impact of proper name CLIR due to the uncertainty of transliteration. The CLIR experiments were conducted using the NTCIR7 IR4QA dataset. The corpus was first corrupted by introducing 4 to 5 different transliterations. These new proper names severely degraded the IR performance. The CLIR performance was recovered by our proposed method.

## References

1. Larkey, L., AbdulJaleel, N., Connell, M.: What's in a Name?: Proper Names in Arabic Cross Language Information Retrieval. CIIR Technical Report, IR-278, Univ. of Amherst (2003)
2. Darwish, K., Doermann, D., Jones, R., Oard, D., Rautiainen, M.: TREC-10 Experiments at University of Maryland CLIR and Video. In: 10th TREC, pp. 549–561 (2002)

3. Meng, H., Chen, B., Lo, W.K., Tang, K.: Generating Phonetic Cognates to Handle Named Entities in English-Chinese Cross-Language Spoken Document Retrieval. In: IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 311–314 (2001)
4. Virga, P., Khudanpur, S.: Transliteration of Proper Names in Cross-lingual Information Retrieval. In: ACL Workshop on Multilingual and Mixed-Language Named Entity Recognition, pp. 57–64 (2003)
5. Bellaachia, A., Amor-Tijani, G.: Proper Nouns in English–Arabic Cross Language Information Retrieval. J. American Society for Information Science and Technology 59(12), 1925–1935 (2008)
6. Chen, H.: -S., Huang, S. -J, Ding, Y. -W., Tasi, S. C.: Proper Name Translation in Cross-Language Information Retrieval. In: 17th COLING-ACL 1998, pp. 232–235 (1998)
7. Kishida, K.: Technical Issues of Cross-Language Information Retrieval: A Review. Information Processing & Management 41(3), 433–455 (2005)
8. Xu, J., Weischedel, R., Nguyen, C.: Evaluating a Probabilistic Model for Cross-Lingual Information Retrieval. In: 24th ACM SIGIR, pp. 105–110 (2001)
9. Kraaij, W., Pohlmann, R., Hiemstra, D.: Twenty-one at TREC-8: Using Language Technology for Information Retrieval. In: 8th TREC, pp. 285–300 (2000)
10. Lavrenko, V., Choquette, M., Croft, W.B.: Cross-lingual relevance models. In: 25th ACM SIGIR, pp. 175–182 (2002)
11. Knight, K., Graehl, J.: Machine Transliteration. Computational Linguistics 24(4), 509–612 (1997)
12. Brown, P.E., Pietra, S.A.D., Mercer, R.L.: The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics 19(2), 263–311 (1993)
13. Gao, W., Wong, K.F., Lam, W.: Improving Transliteration with Precise Alignment of Phoneme Chunks and Using Context Features. In: Myaeng, S.-H., Zhou, M., Wong, K.-F., Zhang, H.-J. (eds.) AIRS 2004. LNCS, vol. 3411, pp. 106–117. Springer, Heidelberg (2005)
14. Li, H.Z., Zhang, M., Su, J.: A Joint Source-Channel Model for Machine Transliteration. In: 42nd ACL, pp. 159–166 (2004)
15. Kumaran, A., Kellner, T.: A Generic Framework for Machine Transliteration. In: 30th ACM SIGIR, pp. 721–722 (2008)
16. Klementiev, A., Roth, D.: Weakly Supervised Named Entity Transliteration and Discovery from Multi-lingual Comparable Corpora. In: 44th ACL, pp. 817–824 (2006)
17. Jiang, L., Zhou, M., Chien, L.F., Niu, C.: Named Entity Translation with Web Mining and Transliteration. In: 20th ICJAI, pp. 1629–1634 (2007)
18. Ponte, J.M., Croft, W.B.: A Language Modeling Approach to Information Retrieval. In: 10th ACM SIGIR, pp. 275–281 (1998)
19. Berger, A., Lafferty, J.: Information Retrieval as Statistical Translation. In: 22nd ACM SIGIR, pp. 222–229 (1999)
20. Sakai, T., Kando, N., Lin, C.J., Mitamura, T., Shima, H., Ji, D., Chen, K.H., Nyberg, E.: Overview of the NTCIR-7 ACLIA IR4QA Task. In: NTCIR-7 Workshop Meeting, pp. 77–114 (2008)
21. Zhai, C.X., Lafferty, J.: A Study of Smoothing Methods for Language Models Applied to Information retrieval. ACM Trans. on Information Systems 22(2), 179–214 (2004)
22. Och, F., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29(1), 19–51 (2003)
23. Papeneni, K.A., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: 40th ACL, pp. 311–318 (2001)
24. Jan, E., Ge, N., Lin, S.H., Roukos, S., Sorensen, J.: A Novel Approach to Proper Name Transliteration. Submitted to ISCSLP 2010