Transliteration Alignment

Vladimir Pervouchine, Haizhou Li
Institute for Infocomm Research
A*STAR, Singapore 138632
{vpervouchine,hli}@i2r.a-star.edu.sg

Bo Lin
School of Computer Engineering
NTU, Singapore 639798
linbo@pmail.ntu.edu.sg

Abstract

This paper studies transliteration alignment, its evaluation metrics and applications. We propose a new evaluation metric, *alignment entropy*, grounded on the information theory, to evaluate the alignment quality without the need for the *gold standard* reference and compare the metric with F-score. We study the use of phonological features and affinity statistics for transliteration alignment at phoneme and grapheme levels. The experiments show that better alignment consistently leads to more accurate transliteration. In transliteration modeling application, we achieve a mean reciprocal rate (MRR) of 0.773 on Xinhua personal name corpus, a significant improvement over other reported results on the same corpus. In transliteration validation application, we achieve 4.48% equal error rate on a large LDC corpus.

1 Introduction

Transliteration is a process of rewriting a word from a source language to a target language in a different writing system using the word's phonological equivalent. The word and its transliteration form a *transliteration pair*. Many efforts have been devoted to two areas of studies where there is a need to establish the correspondence between graphemes or phonemes between a transliteration pair, also known as *transliteration alignment*.

One area is the generative transliteration modeling (Knight and Graehl, 1998), which studies how to convert a word from one language to another using statistical models. Since the models are trained on an aligned parallel corpus, the resulting statistical models can only be as good as the alignment of the corpus. Another area is the transliteration validation, which studies the ways to validate transliteration pairs. For example Knight and Graehl (1998) use the lexicon frequency, Qu and Grefenstette (2004) use the statistics in a monolingual corpus and the Web, Kuo et al. (2007) use probabilities estimated from the transliteration model to validate transliteration candidates. In this paper, we propose using the alignment distance between the a bilingual pair of words to establish the evidence of transliteration candidacy. An example of transliteration pair alignment is shown in Figure 1.

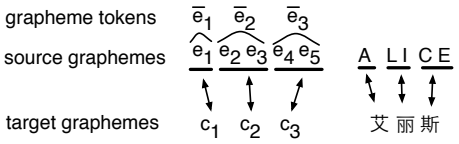


Figure 1: An example of grapheme alignment (Alice, 艾丽斯), where a Chinese grapheme, a character, is aligned to an English grapheme token.

Like the word alignment in statistical machine translation (MT), transliteration alignment becomes one of the important topics in machine transliteration, which has several unique challenges. Firstly, the grapheme sequence in a word is not delimited into grapheme tokens, resulting in an additional level of complexity. Secondly, to maintain the phonological equivalence, the alignment has to make sense at both grapheme and phoneme levels of the source and target languages. This paper reports progress in our ongoing spoken language translation project, where we are interested in the alignment problem of personal name transliteration from English to Chinese.

This paper is organized as follows. In Section 2, we discuss the prior work. In Section 3, we introduce both statistically and phonologically motivated alignment techniques and in Section 4 we advocate an evaluation metric, *alignment entropy* that measures the alignment quality. We report the experiments in Section 5. Finally, we conclude in Section 6.

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 136–144,
Suntec, Singapore, 2-7 August 2009.

2 Related Work

A number of transliteration studies have touched on the alignment issue as a part of the transliteration modeling process, where alignment is needed at levels of graphemes and phonemes. In their seminal paper Knight and Graehl (1998) described a transliteration approach that transfers the grapheme representation of a word via the phonetic representation, which is known as phoneme-based transliteration technique (Virga and Khudanpur, 2003; Meng et al., 2001; Jung et al., 2000; Gao et al., 2004). Another technique is to directly transfer the grapheme, known as direct orthographic mapping, that was shown to be simple and effective (Li et al., 2004). Some other approaches that use both source graphemes and phonemes were also reported with good performance (Oh and Choi, 2002; Al-Onaizan and Knight, 2002; Bilac and Tanaka, 2004).

To align a bilingual training corpus, some take a phonological approach, in which the crafted mapping rules encode the prior linguistic knowledge about the source and target languages directly into the system (Wan and Verspoor, 1998; Meng et al., 2001; Jiang et al., 2007; Xu et al., 2006). Others adopt a statistical approach, in which the affinity between phonemes or graphemes is learned from the corpus (Gao et al., 2004; AbdulJaleel and Larkey, 2003; Virga and Khudanpur, 2003).

In the phoneme-based technique where an intermediate level of phonetic representation is used as the pivot, alignment between graphemes and phonemes of the source and target words is needed (Oh and Choi, 2005). If source and target languages have different phoneme sets, alignment between the the different phonemes is also required (Knight and Graehl, 1998). Although the direct orthographic mapping approach advocates a direct transfer of grapheme at run-time, we still need to establish the grapheme correspondence at the model training stage, when phoneme level alignment can help.

It is apparent that the quality of transliteration alignment of a training corpus has a significant impact on the resulting transliteration model and its performance. Although there are many studies of evaluation metrics of word alignment for MT (Lambert, 2008), there has been much less reported work on evaluation metrics of transliteration alignment. In MT, the quality of training corpus alignment $\mathcal{A}$ is often measured relatively to the *gold standard*, or the ground truth alignment $\mathcal{G}$, which is a manual alignment of the corpus or a part of it. Three evaluation metrics are used: *precision*, *recall*, and F*-score*, the latter being a function of the former two. They indicate how close the alignment under investigation is to the gold standard alignment (Mihalcea and Pedersen, 2003). Denoting the number of cross-lingual mappings that are common in both $\mathcal{A}$ and $\mathcal{G}$ as C_{AG}, the number of cross-lingual mappings in $\mathcal{A}$ as C_A and the number of cross-lingual mappings in $\mathcal{G}$ as C_G, precision Pr is given as C_{AG}/C_A, recall Rc as C_{AG}/C_G and F-score as $2Pr \cdot Rc/(Pr + Rc)$.

Note that these metrics hinge on the availability of the gold standard, which is often not available. In this paper we propose a novel evaluation metric for transliteration alignment grounded on the information theory. One important property of this metric is that it does not require a gold standard alignment as a reference. We will also show that how this metric is used in generative transliteration modeling and transliteration validation.

3 Transliteration alignment techniques

We assume in this paper that the source language is English and the target language is Chinese, although the technique is not restricted to English-Chinese alignment.

Let a word in the source language (English) be $\{e_i\} = \{e_1 \dots e_I\}$ and its transliteration in the target language (Chinese) be $\{c_j\} = \{c_1 \dots c_J\}$, $e_i \in E$, $c_j \in C$, and E, C being the English and Chinese sets of characters, or graphemes, respectively. Aligning $\{e_i\}$ and $\{c_j\}$ means for each target grapheme token $\bar{c}_j$ finding a source grapheme token $\bar{e}_m$, which is an English substring in $\{e_i\}$ that corresponds to c_j, as shown in the example in Figure 1. As Chinese is syllabic, we use a Chinese character c_j as the target grapheme token.

3.1 Grapheme affinity alignment

Given a *distance function* between graphemes of the source and target languages $d(e_i, c_j)$, the problem of alignment can be formulated as a dynamic programming problem with the following function to minimize:

$$\begin{aligned} D_{ij} = min(&D_{i-1,j-1} + d(e_i, c_j), \\ &D_{i,j-1} + d(*, c_j), \\ &D_{i-1,j} + d(e_i, *)) \end{aligned} \quad (1)$$

Here the asterisk * denotes a null grapheme that is introduced to facilitate the alignment between graphemes of different lengths. The minimum distance achieved is then given by

$$D = \sum_{i=1}^{I} d(e_i, c_{\theta(i)}) \quad (2)$$

where $j = \theta(i)$ is the correspondence between the source and target graphemes. The alignment can be performed via the Expectation-Maximization (EM) by starting with a random initial alignment and calculating the *affinity matrix* $count(e_i, c_j)$ over the whole parallel corpus, where element (i, j) is the number of times character e_i was aligned to c_j. From the affinity matrix conditional probabilities $P(e_i|c_j)$ can be estimated as

$$P(e_i|c_j) = count(e_i, c_j) / \sum_j count(e_i, c_j) \quad (3)$$

Alignment $j = \theta(i)$ between $\{e_i\}$ and $\{c_j\}$ that maximizes probability

$$P = \prod_i P(c_{\theta(i)}|e_i) \quad (4)$$

is also the same alignment that minimizes alignment distance D:

$$D = -\log P = -\sum_i \log P(c_{\theta(i)}|e_i) \quad (5)$$

In other words, equations (2) and (5) are the same when we have the distance function $d(e_i, c_j) = -\log P(c_j|e_i)$. Minimizing the overall distance over a training corpus, we conduct EM iterations until the convergence is achieved.

This technique solely relies on the affinity statistics derived from training corpus, thus is called grapheme affinity alignment. It is also equally applicable for alignment between a pair of symbol sequences representing either graphemes or phonemes. (Gao et al., 2004; AbdulJaleel and Larkey, 2003; Virga and Khudanpur, 2003).

3.2 Grapheme alignment via phonemes

Transliteration is about finding phonological equivalent. It is therefore a natural choice to use the phonetic representation as the pivot. It is common though that the sound inventory differs from one language to another, resulting in different phonetic representations for source and target words. Continuing with the earlier example,

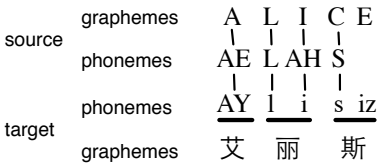


Figure 2: An example of English-Chinese transliteration alignment via phonetic representations.

Figure 2 shows the correspondence between the graphemes and phonemes of English word "Alice" and its Chinese transliteration, with CMU phoneme set used for English (Chase, 1997) and IIR phoneme set for Chinese (Li et al., 2007a).

A Chinese character is often mapped to a unique sequence of Chinese phonemes. Therefore, if we align English characters $\{e_i\}$ and Chinese phonemes $\{cp_k\}$ ($cp_k \in CP$ set of Chinese phonemes) well, we almost succeed in aligning English and Chinese grapheme tokens. Alignment between $\{e_i\}$ and $\{cp_k\}$ becomes the main task in this paper.

3.2.1 Phoneme affinity alignment

Let the phonetic transcription of English word $\{e_i\}$ be $\{ep_n\}$, $ep_n \in EP$, where EP is the set of English phonemes. Alignment between $\{e_i\}$ and $\{ep_n\}$, as well as between $\{ep_n\}$ and $\{cp_k\}$ can be performed via EM as described above. We estimate conditional probability of Chinese phoneme cp_k after observing English character e_i as

$$P(cp_k|e_i) = \sum_{\{ep_n\}} P(cp_k|ep_n) P(ep_n|e_i) \quad (6)$$

We use the distance function between English graphemes and Chinese phonemes $d(e_i, cp_k) = -\log P(cp_k|e_i)$ to perform the initial alignment between $\{e_i\}$ and $\{cp_k\}$ via dynamic programming, followed by the EM iterations until convergence. The estimates for $P(cp_k|ep_n)$ and $P(ep_n|e_i)$ are obtained from the affinity matrices: the former from the alignment of English and Chinese phonetic representations, the latter from the alignment of English words and their phonetic representations.

3.2.2 Phonological alignment

Alignment between the phonetic representations of source and target words can also be achieved using the linguistic knowledge of phonetic similarity. Oh and Choi (2002) define classes of

phonemes and assign various distances between phonemes of different classes. In contrast, we make use of phonological descriptors to define the similarity between phonemes in this paper.

Perhaps the most common way to measure the phonetic similarity is to compute the distances between phoneme features (Kessler, 2005). Such features have been introduced in many ways, such as perceptual attributes or articulatory attributes. Recently, Tao et al. (2006) and Yoon et al. (2007) have studied the use of phonological features and manually assigned phonological distance to measure the similarity of transliterated words for extracting transliterations from a comparable corpus.

We adopt the binary-valued articulatory attributes as the phonological descriptors, which are used to describe the CMU and IIR phoneme sets for English and Chinese Mandarin respectively. Withgott and Chen (1993) define a feature vector of phonological descriptors for English sounds. We extend the idea by defining a 21-element binary feature vector for each English and Chinese phoneme. Each element of the feature vector represents presence or absence of a phonological descriptor that differentiates various kinds of phonemes, e.g. vowels from consonants, front from back vowels, nasals from fricatives, etc[1].

In this way, a phoneme is described by a feature vector. We express the similarity between two phonemes by the Hamming distance, also called the phonological distance, between the two feature vectors. A difference in one descriptor between two phonemes increases their distance by 1. As the descriptors are chosen to differentiate between sounds, the distance between similar phonemes is low, while that between two very different phonemes, such as a vowel and a consonant, is high. The null phoneme, added to both English and Chinese phoneme sets, has a constant distance to any actual phonemes, which is higher than that between any two actual phonemes.

We use the phonological distance to perform the initial alignment between English and Chinese phonetic representations of words. After that we proceed with recalculation of the distances between phonemes using the affinity matrix as described in Section 3.1 and realign the corpus again. We continue the iterations until convergence is reached. Because of the use of phonological descriptors for the initial alignment, we call this technique the *phonological alignment*.

[1]The complete table of English and Chinese phonemes with their descriptors, as well as the transliteration system demo is available at http://translit.i2r.a-star.edu.sg/demos/transliteration/

4 Transliteration alignment entropy

Having aligned the graphemes between two languages, we want to measure how good the alignment is. Aligning the graphemes means aligning the English substrings, called the source grapheme tokens, to Chinese characters, the target grapheme tokens. Intuitively, the more consistent the mapping is, the better the alignment will be. We can quantify the consistency of alignment via *alignment entropy* grounded on information theory.

Given a corpus of aligned transliteration pairs, we calculate $count(c_j, \bar{e}_m)$, the number of times each Chinese grapheme token (character) c_j is mapped to each English grapheme token $\bar{e}_m$. We use the counts to estimate probabilities

$$P(\bar{e}_m, c_j) = count(c_j, \bar{e}_m) / \sum_{m,j} count(c_j, \bar{e}_m)$$

$$P(\bar{e}_m | c_j) = count(c_j, \bar{e}_m) / \sum_{m} count(c_j, \bar{e}_m)$$

The *alignment entropy* of the transliteration corpus is the weighted average of the entropy values for all Chinese tokens:

$$\begin{aligned} H &= -\sum_j P(c_j) \sum_m P(\bar{e}_m|c_j) \log P(\bar{e}_m|c_j) \\ &= -\sum_{m,j} P(\bar{e}_m, c_j) \log P(\bar{e}_m|c_j) \end{aligned} \tag{7}$$

Alignment entropy indicates the uncertainty of mapping between the English and Chinese tokens resulting from alignment. We expect and will show that this estimate is a good indicator of the alignment quality, and is as effective as the F-score, but without the need for a gold standard reference. A lower alignment entropy suggests that each Chinese token tends to be mapped to fewer distinct English tokens, reflecting better consistency. We expect a good alignment to have a sharp cross-lingual mapping with low alignment entropy.

5 Experiments

We use two transliteration corpora: Xinhua corpus (Xinhua News Agency, 1992) of 37,637 personal name pairs and LDC Chinese-English

named entity list LDC2005T34 (Linguistic Data Consortium, 2005), containing 673,390 personal name pairs. The LDC corpus is referred to as LDC05 for short hereafter. For the results to be comparable with other studies, we follow the same splitting of Xinhua corpus as that in (Li et al., 2007b) having a training and testing set of 34,777 and 2,896 names respectively. In contrast to the well edited Xinhua corpus, LDC05 contains erroneous entries. We have manually verified and corrected around 240,000 pairs to clean up the corpus. As a result, we arrive at a set of 560,768 English-Chinese (EC) pairs that follow the Chinese phonetic rules, and a set of 83,403 English-Japanese Kanji (EJ) pairs, which follow the Japanese phonetic rules, and the rest 29,219 pairs (REST) being labeled as incorrect transliterations. Next we conduct three experiments to study 1) alignment entropy vs. F-score, 2) the impact of alignment quality on transliteration accuracy, and 3) how to validate transliteration using alignment metrics.

5.1 Alignment entropy vs. F-score

As mentioned earlier, for English-Chinese grapheme alignment, the main task is to align English graphemes to Chinese phonemes. Phonetic transcription for the English names in Xinhua corpus are obtained by a grapheme-to-phoneme (G2P) converter (Lenzo, 1997), which generates phoneme sequence without providing the exact correspondence between the graphemes and phonemes. G2P converter is trained on the CMU dictionary (Lenzo, 2008).

We align English grapheme and phonetic representations $e - ep$ with the affinity alignment technique (Section 3.1) in 3 iterations. We further align the English and Chinese phonetic representations $ep - cp$ via both affinity and phonological alignment techniques, by carrying out 6 and 7 iterations respectively. The alignment methods are schematically shown in Figure 3.

To study how alignment entropy varies according to different quality of alignment, we would like to have many different alignment results. We pair the intermediate results from the $e - ep$ and $ep - cp$ alignment iterations (see Figure 3) to form $e - ep - cp$ alignments between English graphemes and Chinese phonemes and let them converge through few more iterations, as shown in Figure 4. In this way, we arrive at a total of 114 phonological and 80 affinity alignments of different quality.

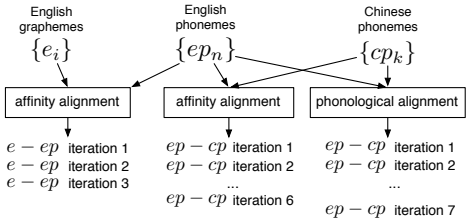


Figure 3: Aligning English graphemes to phonemes $e-ep$ and English phonemes to Chinese phonemes $ep-cp$. Intermediate $e-ep$ and $ep-cp$ alignments are used for producing $e - ep - cp$ alignments.

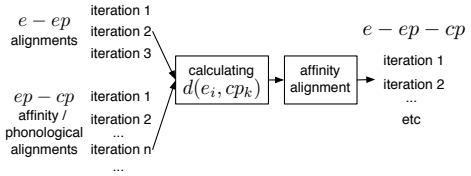
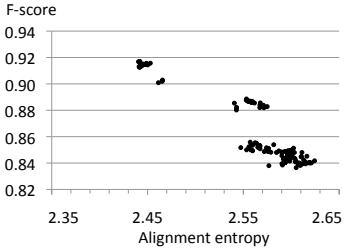


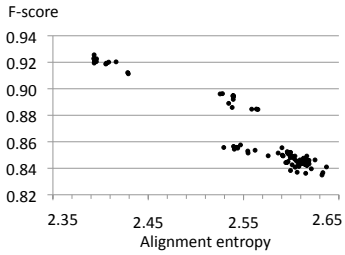
Figure 4: Example of aligning English graphemes to Chinese phonemes. Each combination of $e-ep$ and $ep-cp$ alignments is used to derive the initial distance $d(e_i, cp_k)$, resulting in several $e-ep-cp$ alignments due to the affinity alignment iterations.

We have manually aligned a random set of 3,000 transliteration pairs from the Xinhua training set to serve as the gold standard, on which we calculate the precision, recall and F-score as well as alignment entropy for each alignment. Each alignment is reflected as a data point in Figures 5a and 5b. From the figures, we can observe a clear correlation between the alignment entropy and F-score, that validates the effectiveness of alignment entropy as an evaluation metric. Note that we don't need the gold standard reference for reporting the alignment entropy.

We also notice that the data points seem to form clusters inside which the value of F-score changes insignificantly as the alignment entropy changes. Further investigation reveals that this could be due to the limited number of entries in the gold standard. The 3,000 names in the gold standard are not enough to effectively reflect the change across different alignments. F-score requires a large gold standard which is not always available. In contrast, because the alignment entropy doesn't depend on the gold standard, one can easily report the alignment performance on any unaligned parallel corpus.



(a) 80 affinity alignments



(b) 114 phonological alignments

Figure 5: Correlation between F-score and alignment entropy for Xinhua training set alignments. Results for precision and recall have similar trends .

5.2 Impact of alignment quality on transliteration accuracy

We now further study how the alignment affects the generative transliteration model in the framework of the joint source-channel model (Li et al., 2004). This model performs transliteration by maximizing the joint probability of the source and target names $P(\{e_i\}, \{c_j\})$, where the source and target names are sequences of English and Chinese grapheme tokens. The joint probability is expressed as a chain product of a series of conditional probabilities of token pairs $P(\{e_i\}, \{c_j\}) = P((\bar{e}_k, c_k)|(\bar{e}_{k-1}, c_{k-1}))$, $k = 1 \ldots N$, where we limit the history to one preceding pair, resulting in a bigram model. The conditional probabilities for token pairs are estimated from the aligned training corpus. We use this model because it was shown to be simple yet accurate (Ekbal et al., 2006; Li et al., 2007b). We train a model for each of the 114 phonological alignments and the 80 affinity alignments in Section 5.1 and conduct transliteration experiment on the Xinhua test data.

During transliteration, an input English name is first decoded into a lattice of all possible English and Chinese grapheme token pairs. Then the joint source-channel transliteration model is used to score the lattice to obtain a ranked list of m most likely Chinese transliterations (m-best list).

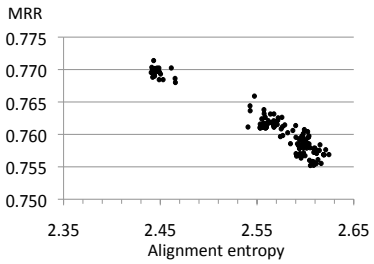
We measure transliteration accuracy as the mean reciprocal rank (MRR) (Kantor and Voorhees, 2000). If there is only one correct Chinese transliteration of the k-th English word and it is found at the r_k-th position in the m-best list, its reciprocal rank is $1/r_k$. If the list contains no correct transliterations, the reciprocal rank is 0. In case of multiple correct transliterations, we take the one that gives the highest reciprocal rank. MRR is the average of the reciprocal ranks across all words in the test set. It is commonly used as a measure of transliteration accuracy, and also allows us to make a direct comparison with other reported work (Li et al., 2007b).

We take $m = 20$ and measure MRR on Xinhua test set for each alignment of Xinhua training set as described in Section 5.1. We report MRR and the alignment entropy in Figures 6a and 7a for the affinity and phonological alignments respectively. The highest MRR we achieve is 0.771 for affinity alignments and 0.773 for phonological alignments. This is a significant improvement over the MRR of 0.708 reported in (Li et al., 2007b) on the same data. We also observe that the phonological alignment technique produces, on average, better alignments than the affinity alignment technique in terms of both the alignment entropy and MRR.

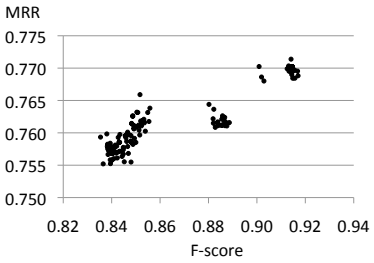
We also report the MRR and F-scores for each alignment in Figures 6b and 7b, from which we observe that alignment entropy has stronger correlation with MRR than F-score does. The Spearman's rank correlation coefficients are -0.89 and -0.88 for data in Figure 6a and 7a respectively. This once again demonstrates the desired property of alignment entropy as an evaluation metric of alignment.

To validate our findings from Xinhua corpus, we further carry out experiments on the EC set of LDC05 containing 560,768 entries. We split the set into 5 almost equal subsets for cross-validation: in each of 5 experiments one subset is used for testing and the remaining ones for training. Since LDC05 contains one-to-many English-Chinese transliteration pairs, we make sure that an English name only appears in one subset.

Note that the EC set of LDC05 contains many names of non-English, and, generally, non-European origin. This makes the G2P converter less accurate, as it is trained on an English phonetic dictionary. We therefore only apply the affinity alignment technique to align the EC set. We

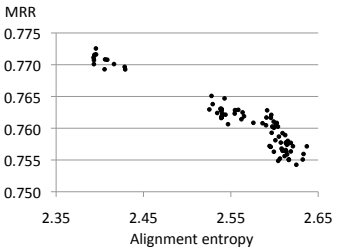


(a) 80 affinity alignments

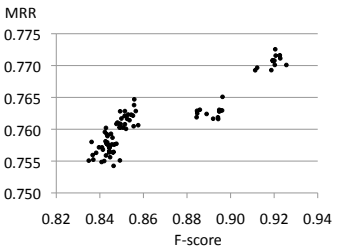


(b) 80 affinity alignments

Figure 6: Mean reciprocal ratio on Xinhua test set vs. alignment entropy and F-score for models trained with different affinity alignments.



(a) 114 phonological alignments



(b) 114 phonological alignments

Figure 7: Mean reciprocal ratio on Xinhua test set vs. alignment entropy and F-score for models trained with different phonological alignments.

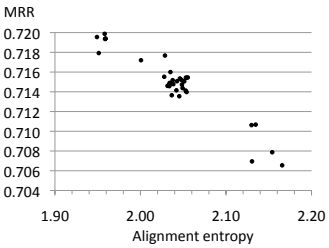


Figure 8: Mean reciprocal ratio vs. alignment entropy for alignments of EC set.

use each iteration of the alignment in the transliteration modeling and present the resulting MRR along with alignment entropy in Figure 8. The MRR results are the averages of five values produced in the five-fold cross-validations.

We observe a clear correlation between the alignment entropy and transliteration accuracy expressed by MRR on LDC05 corpus, similar to that on Xinhua corpus, with the Spearman's rank correlation coefficient of -0.77. We obtain the highest average MRR of 0.720 on the EC set.

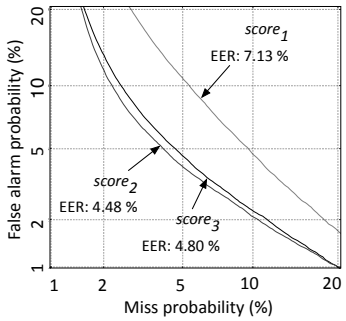
5.3 Validating transliteration using alignment measure

Transliteration validation is a hypothesis test that decides whether a given transliteration pair is genuine or not. Instead of using the lexicon frequency (Knight and Graehl, 1998) or Web statistics (Qu and Grefenstette, 2004), we propose validating transliteration pairs according to the alignment distance D between the aligned English graphemes and Chinese phonemes (see equations (2) and (5)). A distance function $d(e_i, cp_k)$ is established from each alignment on the Xinhua training set as discussed in Section 5.2.

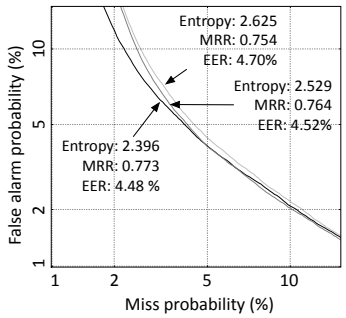
An audit of LDC05 corpus groups the corpus into three sets: an English-Chinese (EC) set of 560,768 samples, an English-Japanese (EJ) set of 83,403 samples and the REST set of 29,219 samples that are not transliteration pairs. We mark the EC name pairs as genuine and the rest 112,622 name pairs that do not follow the Chinese phonetic rules as false transliterations, thus creating the ground truth labels for an English-Chinese transliteration validation experiment. In other words, LDC05 has 560,768 genuine transliteration pairs and 112,622 false ones.

We run one iteration of alignment over LDC05 (both genuine and false) with the distance function $d(e_i, cp_k)$ derived from the affinity matrix of one aligned Xinhua training set. In this way, each transliteration pair in LDC05 provides an alignment distance. One can expect that a genuine transliteration pair typically aligns well, leading to a low distance, while a false transliteration pair will do otherwise. To remove the effect of word length, we normalize the distance by the English name length, the Chinese phonetic transcription

length, and the sum of both, producing $score_1$, $score_2$ and $score_3$ respectively.



(a) DET with $score_1$, $score_2$, $score_3$.



(b) DET results vs. three different alignment quality.

Figure 9: Detection error tradeoff (DET) curves for transliteration validation on LDC05.

We can now classify each LDC05 name pair as genuine or false by having a hypothesis test. When the test score is lower than a pre-set threshold, the name pair is accepted as genuine, otherwise false. In this way, each pre-set threshold will present two types of errors, a false alarm and a miss-detect rate. A common way to present such results is via the detection error tradeoff (DET) curves, which show all possible decision points, and the equal error rate (EER), when false alarm and miss-detect rates are equal.

Figure 9a shows three DET curves based on $score_1$, $score_2$ and $score_3$ respectively for one one alignment solution on the Xinhua training set. The horizontal axis is the probability of miss-detecting a genuine transliteration, while the vertical one is the probability of false-alarms. It is clear that out of the three, $score_2$ gives the best results.

We select the alignments of Xinhua training set that produce the highest and the lowest MRR. We also randomly select three other alignments that produce different MRR values from the pool of 114 phonological and 80 affinity alignments.

Xinhua train set alignment	Alignment entropy of Xinhua train set	MRR on Xinhua test set	LDC classification EER, %
1	2.396	0.773	4.48
2	2.529	0.764	4.52
3	2.586	0.761	4.51
4	2.621	0.757	4.71
5	2.625	0.754	4.70

Table 1: Equal error ratio of LDC transliteration pair validation for different alignments of Xinhua training set.

We use each alignment to derive distance function $d(e_i, cp_k)$. Table 1 shows the EER of LDC05 validation using $score_2$, along with the alignment entropy of the Xinhua training set that derives $d(e_i, cp_k)$, and the MRR on Xinhua test set in the generative transliteration experiment (see Section 5.2) for all 5 alignments. To avoid cluttering Figure 9b, we show the DET curves for alignments 1, 2 and 5 only. We observe that distance function derived from better aligned Xinhua corpus, as measured by both our alignment entropy metric and MRR, leads to a higher validation accuracy consistently on LDC05.

6 Conclusions

We conclude that the alignment entropy is a reliable indicator of the alignment quality, as confirmed by our experiments on both Xinhua and LDC corpora. Alignment entropy does not require the gold standard reference, it thus can be used to evaluate alignments of large transliteration corpora and is possibly to give more reliable estimate of alignment quality than the F-score metric as shown in our transliteration experiment.

The alignment quality of training corpus has a significant impact on the transliteration models. We achieve the highest MRR of 0.773 on Xinhua corpus with phonological alignment technique, which represents a significant performance gain over other reported results. Phonological alignment outperforms affinity alignment on clean database.

We propose using alignment distance to validate transliterations. A high quality alignment on a small verified corpus such as Xinhua can be effectively used to validate a large noisy corpus, such as LDC05. We believe that this property would be useful in transliteration extraction, cross-lingual information retrieval applications.

References

Nasreen AbdulJaleel and Leah S. Larkey. 2003. Statistical transliteration for English-Arabic cross language information retrieval. In *Proc. ACM CIKM*.

Yaser Al-Onaizan and Kevin Knight. 2002. Machine transliteration of names in arabic text. In *Proc. ACL Workshop: Computational Apporaches to Semitic Languages*.

Slaven Bilac and Hozumi Tanaka. 2004. A hybrid back-transliteration system for Japanese. In *Proc. COLING*, pages 597–603.

Lin L. Chase. 1997. *Error-responsive feedback mechanisms for speech recognizers*. Ph.D. thesis, CMU.

Asif Ekbal, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2006. A modified joint source-channel model for transliteration. In *Proc. COLING/ACL*, pages 191–198

Wei Gao, Kam-Fai Wong, and Wai Lam. 2004. Phoneme-based transliteration of foreign names for OOV problem. In *Proc. IJCNLP*, pages 374–381.

Long Jiang, Ming Zhou, Lee-Feng Chien, and Cheng Niu. 2007. Named entity translation with web mining and transliteration. In *IJCAI*, pages 1629–1634.

Sung Young Jung, SungLim Hong, and Eunok Paek. 2000. An English to Korean transliteration model of extended Markov window. In *Proc. COLING*, volume 1.

Paul. B. Kantor and Ellen. M. Voorhees. 2000. The TREC-5 confusion track: comparing retrieval methods for scanned text. *Information Retrieval*, 2:165–176.

Brett Kessler. 2005. Phonetic comparison algorithms. *Transactions of the Philological Society*, 103(2):243–260.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4).

Jin-Shea Kuo, Haizhou Li, and Ying-Kuei Yang. 2007. A phonetic similarity model for automatic extraction of transliteration pairs. *ACM Trans. Asian Language Information Processing*, 6(2).

Patrik Lambert. 2008. *Exploiting lexical information and discriminative alignment training in statistical machine translation*. Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona, Spain.

Kevin Lenzo. 1997. t2p: text-to-phoneme converter builder. http://www.cs.cmu.edu/~lenzo/t2p/.

Kevin Lenzo. 2008. The CMU pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Haizhou Li, Min Zhang, and Jian Su. 2004. A joint source-channel model for machine transliteration. In *Proc. ACL*, pages 159–166.

Haizhou Li, Bin Ma, and Chin-Hui Lee. 2007a. A vector space modeling approach to spoken language identification. *IEEE Trans. Acoust., Speech, Signal Process.*, 15(1):271–284.

Haizhou Li, Khe Chai Sim, Jin-Shea Kuo, and Minghui Dong. 2007b. Semantic transliteration of personal names. In *Proc. ACL*, pages 120–127.

Linguistic Data Consortium. 2005. LDC Chinese-English name entity lists LDC2005T34.

Helen M. Meng, Wai-Kit Lo, Berlin Chen, and Karen Tang. 2001. Generate phonetic cognates to handle name entities in English-Chinese cross-language spoken document retrieval. In *Proc. ASRU*.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *Proc. HLT-NAACL*, pages 1–10.

Jong-Hoon Oh and Key-Sun Choi. 2002. An English-Korean transliteration model using pronunciation and contextual rules. In *Proc. COLING 2002*.

Jong-Hoon Oh and Key-Sun Choi. 2005. Machine learning based english-to-korean transliteration using grapheme and phoneme information. *IEICE Trans. Information and Systems*, E88-D(7):1737–1748.

Yan Qu and Gregory Grefenstette. 2004. Finding ideographic representations of Japanese names written in Latin script via language identification and corpus validation. In *Proc. ACL*, pages 183–190.

Tao Tao, Su-Youn Yoon, Andrew Fisterd, Richard Sproat, and ChengXiang Zhai. 2006. Unsupervised named entity transliteration using temporal and phonetic correlation. In *Proc. EMNLP*, pages 250–257.

Paola Virga and Sanjeev Khudanpur. 2003. Transliteration of proper names in cross-lingual information retrieval. In *Proc. ACL MLNER*.

Stephen Wan and Cornelia Maria Verspoor. 1998. Automatic English-Chinese name transliteration for development of multilingual resources. In *Proc. COLING*, pages 1352–1356.

M. M. Withgott and F. R. Chen. 1993. *Computational models of American speech*. Centre for the study of language and information.

Xinhua News Agency. 1992. *Chinese transliteration of foreign personal names*. The Commercial Press.

LiLi Xu, Atsushi Fujii, and Tetsuya Ishikawa. 2006. Modeling impression in probabilistic transliteration into Chinese. In *Proc. EMNLP*, pages 242–249.

Su-Youn Yoon, Kyoung-Young Kim, and Richard Sproat. 2007. Multilingual transliteration using feature based phonetic method. In *Proc. ACL*, pages 112–119.