Chapter 7: Experiments on Transliteration Alignment

In this chapter, we describe the setup for the experiments and discuss the performance evaluation of the proposed transliteration model when applied to align bilingual transliteration pairs in parallel corpora.

7.1 Experimental Setup

Several corpora were collected to estimate the parameters of the proposed models and to evaluate the performance of the proposed approach. The corpus *T0* for training consisted of 2,430 pairs of English names and transliterations in Chinese. The training corpus, composed of a bilingual proper name list, was collected from "Handbook of English Name Knowledge" edited by Huai (1989). The bilingual proper name list consists of first names, last names, and nicknames. For example, (Adolf, 阿道夫 "Ataofu") and (Adelaide, 阿德萊德 "Atelaite") are first names, (Abbey, 阿比 "Api") and (Adela, 阿德拉 "Atela") are last names, and (Archie, 阿爾奇 "Aerhchi")

and (Allie, 阿莉 “Ali”) are nicknames, for males and females, respectively. Some first names are also used as last names. For instance, “Abel” can be either a first name or a last name. Table 7.1 shows some examples of the training corpus.

Table 7.1 Some samples from the training set T0.

Source word	Target word	Source word	Target word
Abe	阿貝	Agatha	阿佳莎
Abbey	阿比	Acton	阿克頓
Abbot	阿伯特	Arkwright	阿克賴特
Archer	阿徹	Arabella	阿拉蓓拉
Adolf	阿道夫	Alaric	阿拉里克
Adolphus	阿道弗斯	Alasdair	阿拉斯代爾
Adela	阿德拉	Alastair	阿拉斯泰爾
Adelaide	阿德萊德	Alethea	阿蕾西
Arden	阿登	Alonzo	阿朗索
Albert	阿爾伯特	Ariadne	阿莉雅德妮
Alfonso	阿爾方索	Allegra	阿莉葛拉
Alfie	阿爾菲	Alister	阿利斯特
Alf	阿爾夫	Allie	阿莉
Algy	阿爾吉	Arlene	阿琳
Algernon	阿爾傑農	Alan	阿倫
Alma	阿爾瑪	Aloys	阿洛伊斯
Almeric	阿爾梅里克	Aloysius	阿洛伊修斯
Archie	阿爾奇	Amadeus	阿瑪迪斯
Alva	阿爾娃	Amabel	阿瑪蓓兒
Alphonsus	阿方薩斯	Amanda	阿曼妲
Alphonso	阿方索	Amelia	阿蜜莉雅
Afra	阿芙拉	Arms	阿姆斯
Avril	阿弗里爾	Armstrong	阿姆斯特朗
Agnes	阿莨妮絲	Anastasia	阿娜絲塔西雅
Argus	阿格斯	Arno	阿諾

In the experiment, three sets of parallel-aligned texts (Chuang et al., 2002), *P1*, *P2*, and *P3*, were prepared to evaluate the performance of proposed methods. *P1* consisted of 500 bilingual examples from the English-Chinese version of the Longman Dictionary of Contempory English (LDOCE) (Proctor, 1988). *P2* consisted of 300 aligned sentences from *Scientific American*, USA and Taiwan Editions[6]. *P3* consisted of 300 aligned sentences from the *Sinorama* Corpus.

In the experiment, we dealt with person and place names as well as their transliterations from the parallel corpora. The performance of transliteration extraction was evaluated based on the precision rates of transliteration words or characters. For simplicity, we considered each proper name in the source sentence in turn and determined its corresponding transliteration independently. Table 7.2 shows some examples from the testing set *P1*.

7.2 TUs for English and Chinese

The proposed model is based on TUs, which are more linguistically motivated than individual characters. Table 7.3 lists some of the most frequently occurring English TUs of length 1 to 3. Table 7.4 lists some of the most frequently occurring Chinese

[6] *Scientific American*: "http://www.sciam.com" (USA edition) and "http://www.sciam.com.tw" (Taiwan edition).

TUs. Table 7.5 shows some English-Chinese TU-mapping probabilities automatically estimated from all of the training data.

Table 7.2 Some bilingual examples from the testing set *P1*.

He is a (second) Caesar in speech and leadership. 他在演說及領導方面的才能有如凱撒再世.
Hamlet kills the king in Act 5 Scene 2. 哈姆雷特在第五幕第二景中把國王殺死.
Can you adduce any reason at all for his strange behaviour, Holmes? 福爾摩斯，你能否舉出什麼理由解釋他的古怪行為?
To see George, of all people, in the Ritz Hotel! 真想不到，居然在麗澤旅館看到喬治!
He has 2 caps for playing cricket for England. 他代表英國打板球而得到兩頂榮譽帽.
They appointed him to catch all the rats in Hamelin. 他們指派他捉漢姆林區所有的老鼠.
Burlington Arcade is a famous shopping passage in London. 柏靈頓拱廊是倫敦有名的購物街.
The architecture of ancient Greece. 古希臘的建築風格.
Drink Rossignol, the aristocrat of table wines! 喝羅西諾酒吧！這是餐酒中的上品!
Cleopatra was bitten by an asp. 克利奧佩特拉女王是被小毒蛇咬死的.
I shall soon be leaving for an assignment in India. 我很快就要去印度擔任一項職務.
Our plane stopped at London (airport) on its way to New York. 我們的飛機往紐約，途中在倫敦機場過境.
Schoenberg used atonality in the music of his middle period. 桑伯格在中期用無調性方式作曲.
This tune is usually attributed to J. S. Bach. 這個曲子通常被認為是巴哈所作.
Now that this painting has been authenticated as a Rembrandt, it's worth 10 times as much as I paid for it! 由於這幅畫已證實是倫布朗真蹟，它的時價是我當初買下來時的十倍!

Table 7.3 Some high frequency English TUs.

Length of English TU *u*	High frequency TUs
1	a, e, i, n, l, s, o, r, d, t
2	er, ie, ar, ll, th, or, ch, tt, ck, ph
3	lle, sch

Table 7.4 Some high frequency Chinese TUs.

Length of Chinese TU *v*	High frequency TUs
1	i, a, l, n, o, t, e, p, m, u
2	te, ei, ai, ch, ko, hs, ng, ao, pu, fu
3	ssu, erh, ieh, chi, hsi
4	shih
5	chieh

Table 7.5 English-Chinese TU-mapping probabilities.

u	*v*	$P(v\|u)$
	h	0.272
ae	ei	0.571
ae	i	0.214
ai	a	0.500
ai	e	0.250
ar	a	0.794
au	o	0.772
aw	ao	0.545
aw	o	0.454

u	*v*	$P(v\|u)$
ei	i	0.900
eu	yu	0.785
ew	u	0.500
ey	i	0.998
f	f	0.586
ff	f	0.733
ff	fu	0.266
g	ko	0.350
g	ch	0.345

The automatic learning process resulted in mostly regular monographs and digraphs found in pronunciation dictionaries, such as the *Longman Pronunciation Dictionary* (LPD) (Wells, 2001), including “rh” and “au.” However, it also learned additional TUs, such as “cq” in the person names “Jacqueline” and “Jacquetta.” For example, after the second iteration of EM training, the most likely TU alignment sequence of the name pair (Jacqueline, Chiehkueilin “傑桂琳”) is shown in Figure 7.1.

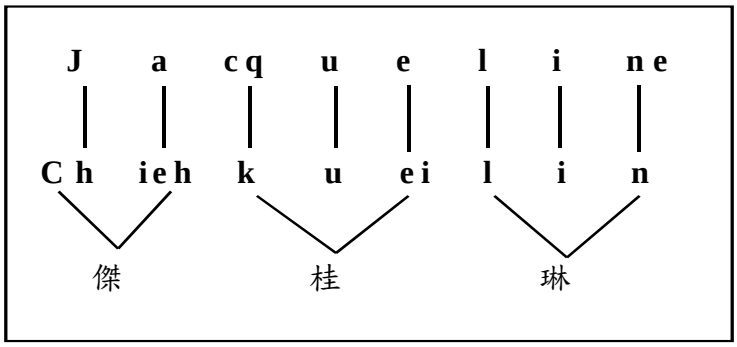


Figure 7.1 TU alignment of the name pair (Jacqueline, Chiehkueilin “傑桂琳”).

It should be noted that an original word may have more than one transliteration. For instance, the English name “Beaufort” has several possible Chinese terms {“鮑福” (Paofu), “鮑佛” (Paofo), “蒲福” (Pufu), “鮑佛特” (Paofote)}. The TUs of the word “Beaufort” were automatically and dynamically constructed and aligned with their corresponding transliteration TUs via the proposed model. The results are shown in Figure 7.2.

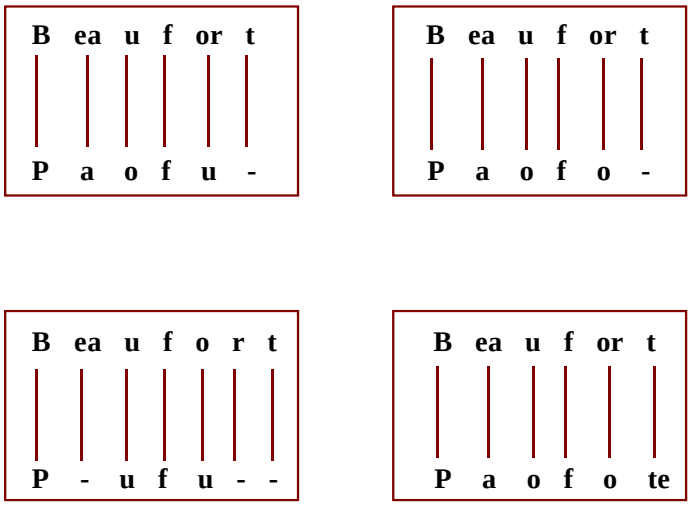
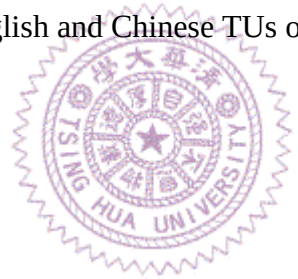


Figure 7.2 TU alignment of “Beaufort” and corresponding transliterations.

Although Knight and Graehl (1998) applied EM to automatically learn similarities of English-Japanese name pairs, English words and Japanese katakana words have to be converted into English sounds and Japanese sounds, respectively, via pronunciation dictionaries. Each English sound can map to one or more Japanese sounds. Compared with their study, one of the advantages of our approach is that we do not have to find the exact pronunciations via dictionary lookup or various grapheme-to-phoneme rules. To be more specific, a set of often-used Chinese characters for transliteration was selected from the collected corpora. Although many Chinese characters have more than one pronunciation, we found that almost all the characters used for transliteration have unique pronunciations. For those Chinese characters not used for transliteration, we choose the most frequently used

pronunciation instead. Since we focus on transliterated words, we do not apply any Chinese pronunciation disambiguation algorithm to decide the exact pronunciation for each character. Thus, the romanization of Chinese Characters can be conducted directly via table lookup instead of using a pronunciation dictionary. Moreover, to accelerate the convergence of EM training and reduce noisy TU pairs at grapheme-level string mapping, we adopt a many-to-many mapping under the constraints of a limited set of matched types based on phonetic knowledge. The maximum lengths of English and Chinese TUs are 3 and 5, respectively. Table 7.6 shows the match types and English and Chinese TUs obtained in our experiments.



7.3 Evaluation Metric

In the experiment, the performance of transliteration extraction was evaluated based on precision and recall rates at the word and character levels. Since we considered exactly one proper name in the source language and one transliteration in the target language at a time, the word recall rates were same as the word precision rates:

$$Word\ \ Precision\ \ (WP) = \frac{number\ \ of\ \ correctly\ \ extracted\ \ words}{number\ \ of\ \ correct\ \ words}. \qquad (7.1)$$

The character level recall and precision rates were defined as follows:

$$Character\ \ \ precision\ \ (CP) = \frac{number\ \ of\ \ correctly\ \ extracted\ \ characters}{number\ \ of\ \ extracted\ \ characters}, \qquad (7.2)$$

$$Character\ \ Recall\ (CR) = \frac{number\ of\ correctly\ extracted\ characters}{number\ of\ correct\ characters}. \qquad (7.3)$$

Table 7.6 Examples for each match type.

Match type	TU pair
0 – 1	(, h), (, i), (, n), (, u)
1 – 0	(h,), (k,), (d,), (t,)
1 – 1	(r, l), (y, i), (m, m)
1 – 2	(j, ch), (f, fu), (d, te)
1 – 3	(s, ssu), (l, erh), (r, erh)
1 – 4	(s, shih)
2 – 0	(gh,)
2 – 1	(bb, p), (ey, i), (mm, m)
2 – 2	(dg, ch), (wh, hu), (ck, ko)
2 – 3	(le, erh), (re, erh), (ce, ssu)
2 – 4	(ce, shih)
2 – 5	(ge, chieh)
3 – 2	(sch, hs)
3 – 3	(lle, erh)
3 – 4	(sch, shih)

7.4 Experimental Results and Discussion

In the experiment of extracting transliterations on the data set *P1*, the TM model achieved, on average, a word precision rate of 86%, a character precision rate of

94.4%, and a character recall rate of 96.3%, as shown in Table 7.7. The performance could be further improved by means of simple statistical and linguistic processing, as shown in Table 7.7.

Table 7.7 The experimental results of transliterated word extraction.

Test Set	Methods	WP	CP	CR
P1 (*LDOCE*)	TM	86.0%	94.4%	96.3%
	TM+R1	88.6%	95.4%	97.7%
	TM+R2	90.8%	97.4%	95.9%
	TM+R1+R2	94.2%	98.3%	97.7%
P2 (*Scientific American*)	TM	90.7%	96.9%	97.3%
	TM+R1	92.7%	97.6%	97.9%
	TM+R2	92.0%	97.8%	97.3%
	TM+R1+R2	94.0%	98.3%	97.9%
P3 (*Sinorama*)	TM	86.7%	94.2%	96.1%
	TM+R1	89.0%	94.9%	96.8%
	TM+R2	87.7%	95.8%	94.9%
	TM+R1+R2	93.0%	96.5%	96.7%

Table 7.8 shows some examples of Chinese transliterated words, correctly extracted using the TM model, from *P1*. Although, the TM model failed in some cases, most of these problems could be overcome through the addition of simple linguistic processing, as shown in Table 7.9. The error in the case of “Quirk” occurred because “Quirk” is much closer to “克和格 (kohoko)” than to “柯克 (KoKo),” based on phonetic similarity. In this case, the Chinese transliteration plainly cannot be

correctly extracted. Similar problems, due to similarities at the grapheme level, occurred with the name pairs (Tom, 湯姆 "tangmu") and (John, 約翰 "yuehhan"), as shown in Table 7.9. It is obvious that a collection of commonly used or highly varying transliterations can be incrementally added to a lookup list to further improve the system performance.

We have also performed the same experiments on the data sets *P2* and *P3*, and the results are shown in Table 7.7. Although the performance of the TM approach on the data sets *P1* and *P3* are worse than that of *P2*, obviously, the integrated scheme (TM+R1+R2) exhibits considerable robustness in extracting transliterated words from different data sets in various domains. The results in Table 7.10 show the average rates of word and character precision for the test sets are around 93.8% and 97.8%, respectively.

Table 7.8 Some examples of Chinese transliterations, correctly extracted by the TM model, from *P1*.

Bilingual Sentence	Baseline
He is a second Caesar in speech and leadership. 他在演說及領導方面的才能有如凱撒再世.	凱撒 (kaisa)
In this case I'm acting for my friend Mr. Smith. 我代表我的朋友史密斯先生處理此事.	史密斯 (shihmissu)
What's your alibi for being late this time Jones? 仲斯, 你這次遲到又有什麼藉口?	仲斯 (chungssu)
Can you adduce any reason at all for his strange behaviour, Holmes? 福爾摩斯, 你能否舉出什麼理由解釋他的古怪行為?	福爾摩斯 (fuerhmossu)
They appointed him to catch all the rats in Hamelin. 他們指派他捉漢姆林區所有的老鼠.	漢姆林 (hanmulin)
Drink Rossignol, the aristocrat of table wines! 喝羅西諾酒吧! 這是餐酒中的上品!	羅西諾 (lohsino)
Cleopatra was bitten by an asp. 克利奧佩特拉女王是被小毒蛇咬死的.	克利奧佩特拉 (koliaopeitela)
Schoenberg used atonality in the music of his middle period. 桑伯格在中期用無調性方式作曲.	桑伯格 (sangpoko)
If you have to change trains in London, you may be able to book through to your last station. 假如你要在倫敦換火車的話, 你可以買一張分段乘車到最後一站的車票.	倫敦 (luntun)
This tune is usually attributed to J. S. Bach. 這個曲子通常被認為是巴哈所作.	巴哈 (paha)
Byron awoke one morning to find himself famous. 拜倫一朝醒來發現自己已經成名.	拜倫 (pailun)
You must have kissed the Blarney Stone to be able to talk like that! 你一定是吻過布拉尼的石頭才能夠把話講得那樣動聽!	布拉尼 (pulani)
Quirk and Greenbaum collaborated on the new grammar. 柯克和格林邦合著這本新文法的書.	格林邦 (kolinpang)

Table 7.9 Some examples of possible Chinese transliterations extracted by the proposed approaches.

("*" means the Chinese transliterated words are not correctly extracted.)

Bilingual Sentence	Baseline	+R1	+R2	+R1+R2
David, as you know, writes dictionaries. 你也知道, 大衛的工作是編寫詞典.	大衛的 (taweite)	大衛的	大衛 (tawei)	大衛
The Mediterranean Sea bathes the sunny shores of Italy. 地中海沿著陽光普照的意大利海岸流動著.	的意大利海岸 (teitalihaian)	地中海 (tichunghai)	的意大利海岸	地中海
You have borne yourself bravely in this battle, Lord Faulconbridge. 佛肯伯里爵士, 在這場戰役你表現甚為英勇!	佛肯伯里爵 (fokenpolichueh)	佛肯伯里爵	佛肯伯里 (fokenpoli)	佛肯伯里
Ancient Rome and Greece. 古羅馬及希臘.	及希 (chihsi)	希臘	及希	希臘
Jane is blossoming out into a beautiful girl. 珍已長大成為一個漂亮的女孩子.	珍已 (cheni)	珍已	珍 (chen)	珍
Tom likes to boss younger children about. 湯姆喜歡對較年幼的小孩子發號施令.	湯 (tang)	湯	湯	湯*
Quirk and Greenbaum collaborated on the new grammar. 柯克和格林邦合著這本新文法的書.	克和格 (kohoko)	克和格	克和格	克和格*
John seems to have made a real conquest of Janet. They're always together. 約翰好像真的已經贏得珍娜的芳心, 他們常在一起.	珍 (Jen)	珍	珍	珍*

Table 7.10 The average rates of transliterated word extraction for overall corpora.

Methods	WP	CP	CR
TM	87.5%	95.1%	96.6%
TM+R1	89.8%	95.9%	97.5%
TM+R2	90.3%	97.1%	96.1%
TM+R1+R2	93.8%	97.8%	97.5%

Compared with the previous work, the proposed approach has three advantages. First, the proposed method learns the parameters of the model automatically from a list of bilingual name pairs without using a pronunciation dictionary or grapheme-to-phoneme rules for the source words. Second, the proposed framework is easier to port to other language pairs as long as there is some transliteration training data. Third, the proposed approach matches TUs in the two languages directly, therefore accelerates the matching process by skipping the grapheme-to-phoneme phase.