# Evaluating Chinese-English Translation Systems for Personal Name Coverage

**Benjamin K. Tsou and Oi Yee Kwong**

Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
rlbtsou@uxmail.cityu.edu.hk rlolivia@cityu.edu.hk

## Abstract

This paper discusses the challenges which Chinese-English machine translation (MT) systems face in translating personal names. We show that the translation of names between Chinese and English is complicated by different factors, including orthographic, phonetic, geographic and social ones. Four existing systems were tested for their capability in translating personal names from Chinese to English. Test data embodying geographic and sociolinguistic differences were obtained from a synchronous Chinese corpus of news media texts. It is obvious that systems vary considerably in their ability to identify personal names in the source language and render them properly in the target language. Given the criticality of personal name translation to the overall intelligibility of a translated text, the coverage of personal names should be one of the important criteria in the evaluation of MT performance. Moreover, name translation, which calls for a hybrid approach, would remain a central issue to the future development of MT systems, especially for online and real-time applications.

## Keywords

Personal name translation, MT evaluation, Chinese-English MT, Translation systems, Region-sensitive Romanisation

## Introduction

There are many different aspects to consider in the evaluation of machine translation (MT) systems, including intelligibility, accuracy, error analysis and so on (e.g. Arnold et al., 1994; EAGLES, 1999). However, few seem to directly address the identification and translation of personal names within texts, especially between languages of different families, typically Chinese and English. While there are studies on the identification of personal names from Chinese texts (e.g. Sun et al., 1995; Chen & Bai, 1998), this paper discusses the importance of the translation component and evaluates several existing Chinese-English MT systems for their capability in this context.

The translation of personal names might seem trivial between some languages. For instance, Bill Clinton is always Bill Clinton, in English or in French, written and sometimes even pronounced the same. However, between languages from different families such as Chinese and English, the complexity is often beyond description by a few simple rules. For example, the international Kung Fu film star, 成龍, is known all over the world as Jackie Chan, which phonetically resembles neither its Cantonese (Shing Lung) nor Mandarin (Cheng Long) pronunciation.

Consider some hypothetical MT systems (a) to (d), which produce the following translation output respectively:

(a) The cinema is showing the new film starring Jackie Chan as a police officer.

(b) The cinema is showing the new film starring Shing Lung as a police officer.

(c) The cinema is showing the new film starring Cheng Long as a police officer.

(d) The cinema is showing the new film starring become dragon as a police officer.

It is obvious that assuming everything else being the same, the translation of personal names plays an important role in determining the intelligibility and accuracy of the translation output, as only sentence (a) would make the most sense. The output in (b) and (c) might marginally convey the intended meaning to Chinese readers, but barely so to the rest of the world. Example (d), translating a name as isolated characters, is simply unacceptable. In other words, the generally recognised nominal representation of a person in the target language might be more than a simple phonetic transcription of the source, and anything other than the recognised representation is unlikely to make any sense to anyone at all.

Hence, the ability to identify personal names in the source language and to render them properly in the target language could be critical to the overall intelligibility of the translation results, especially when the source and target languages are from different linguistic families. Therefore the coverage of personal names should be one of the most important criteria in the evaluation of MT performance.

In the following, we will first discuss the challenges faced by MT systems in translating personal names between Chinese and English, and how the problem might be handled by MT systems. We then briefly introduce the four translation systems evaluated in this study, and report on how well these systems translated names from Chinese to English. Finally we will conclude with some implications which this study might have on improving Chinese-English translation systems.

## Challenges of Personal Name Translation

The translation of names between Chinese and English is affected by a complex interaction of different factors including orthographic, phonetic, geographic and social ones. The challenge is thus manifold.

In the most general case, a personal name consists of a surname (Family name, last name) and a given name (first name), but the conventional order of the constituents is different for English names and Chinese names. When we talk about translating names from Chinese to English, it could be one of the following three cases:

(1) to put the (already translated) name back to its authentic English form, e.g.:

蓋茨 → Bill Gates
森柏斯 → Pete Sampras;

(2) to put an authentic Chinese name in an English form, e.g.:

江澤民 → Jiang Zemin
張德培 → Michael Chang

(3) to find the corresponding English form for the Chinese form originating from a third language such as Japanese and Korean, e.g.:

酒井法子 → Noriko Sakai
金大中 → Kim Dae Jung

In this paper, we shall discuss the first two types of challenge, which correspond to the direct translation between Chinese and English. Before we explore how an MT system might handle name translation algorithmically and evaluate existing systems in this regard, we first go through some of the common patterns observed in the first two types of name translation from Chinese to English.

## Translated Chinese to Authentic English

To be able to go back to the authentic English name, one must know how the Chinese form was obtained in the first place, which often followed one of the patterns below. While most renditions of non-Chinese names are based on phonetic properties, others consider meaning and stylistic structures making the translation like a real Chinese name.

**1. Translation of only last name**
Examples: David Beckham, Al Gore
.

| *Given* | *Last* |
|---|---|
| **David** | **Beckham** |
| | 碧咸 |

| *Given* | *Last* |
|---|---|
| **Al** | **Gore** |
| | 戈爾 |

**2. Translation of only given name**
Examples: Camilla Parker Bowles, Hillary Clinton

| *Given* | *Last* |
|---|---|
| **Camilla** | **Parker Bowles** |
| 卡米拉 | |

| *Given* | *Last* |
|---|---|
| **Hillary** | **Clinton** |
| 希拉莉 | |

**3. Translation of full name**
Examples: Julia Roberts, Tom Cruise

| *Given* | *Last* |
|---|---|
| **Julia** | **Roberts** |
| 茱莉亞 | 羅拔絲 |

| *Given* | *Last* |
|---|---|
| **Tom** | **Cruise** |
| 湯 | 告魯斯 |

**4. Transformation of whole name**
Examples: Bernard Shaw, Lavender Patten

| *Given* | *Last* |
|---|---|
| **Bernard** | **Shaw** |
| 蕭伯納 | |

| *Given* | *Last* |
|---|---|
| **Lavender** | **Patten** |
| 林穎彤 | |

Thus the formal or widely accepted Chinese translation of an English name may vary not only in the way the translation is done, but also in the resultant form of the translated name. For instance, a Chinese rendition of an English name may have one to as many as six or seven characters, or even more, which is very different from the form of authentic Chinese names. Moreover, it appears that there are different conventions for translating names of different groups, such as politicians, movie stars, sports players, and other celebrities. Despite the variation, most translations, especially those following the first three patterns above, are phonetically based. Hence, although a transliteration of 碧咸 back to Beckham says nothing on his first name, the translation suffices for comprehension, assuming there is only one well-known Beckham in the talks among people.

## Authentic Chinese to Translated English

The most straightforward translation of a Chinese name is by phonetic transcription with Roman letters. However, not all cases comply with this rule. Generally speaking, the translation of Chinese names into English follows one of the patterns below.

**1. Character Romanisation**[1]
Examples: Li Ka-shing (Hong Kong tycoon), Tung Chee Hwa (Chief Executive of Hong Kong)

| *Last* | *Given* |
|---|---|
| 李 | 嘉誠 |
| Li | Ka-shing |

| *Last* | *Given* |
|---|---|
| 董 | 建華 |
| Tung | Chee Hwa |

**2. English given name added to / replacing Chinese**
Examples: Donald Tsang Yam-kuen (Chief Secretary of Hong Kong), Bruce Lee (Chinese Kung Fu master)

| *Last* | *Given* |
|---|---|
| 曾 | 蔭權 |
| Donald Tsang Yam-kuen | |

| *Last* | *Given* |
|---|---|
| 李 | 小龍 |
| Bruce Lee | |

**3. Husband's surname on top of maiden name**
Examples: Anson Chan (former Chief Secretary of Hong Kong), Betty Tung (Mrs. Tung Chee Hwa)

| *Last* | *Given* |
|---|---|
| 陳 方 | 安生 |
| Anson Chan | |

| *Last* | *Given* |
|---|---|
| 董 趙 | 洪娉 |
| Betty Tung | |

[1] The Romanisation for individual names might follow different and idiosyncratic rules due to regional difference, but for each name there is always a legally recognised Romanisation (see Regional Variation section).

Keeping two surnames is relatively unique for Chinese married women, unlike their English or Japanese counterparts whose maiden names are often replaced by husbands' surnames. The English renditions of such Chinese names, however, follow the conventions of English such that the maiden name is usually not mentioned, unless the name is spelled out in full.

**4. Pseudo Chinese names (especially for artistes)**
Examples: Jackie Chan (Hong Kong film star), Elle Choi (Hong Kong pop singer)

| *Given* | *Last* |
|---|---|
| 成龍 | |
| Jackie | Chan |

| *Given* | *Last* |
|---|---|
| 小雪 | |
| Elle | Choi |

Hence we have seen the different translation patterns of names presenting different levels of difficulty for any MT systems. There is no straightforward reversibility such that translations (in either direction) might not always be readily derivable, and therefore MT systems must have enough coverage of personal names to perform the task well. A large resource for personal name translation is especially important for idiosyncratic mappings which there is simply no chance for any MT system to derive algorithmically, such as the pseudo names mentioned above. There would be no solution except acquiring specific knowledge databases from up-to-date sources in order to handle these names properly.

## Regional Variation

Even with Romanisation alone, translation of names could still be complicated by dialectal difference. In the Chinese writing system, the same characters are used in writing mutually unintelligible dialects, e.g. Cantonese and Mandarin. While Mandarin follows a standard Hanyu Pinyin for Romanisation, there are distinctive Romanisation patterns in Hong Kong, Taiwan, and Singapore, among others. For example, the following names with the same Chinese last name (吳) demonstrate this difference:

吳清輝 **Ng** Ching-fai (Hong Kong Legislative Councillor)
吳邦國 **Wu** Bangguo (Chinese Vice Premier)
吳作棟 **Goh** Chok Tong (Prime Minister of Singapore)

Hence we said that personal name translation between Chinese and English involves a complex interaction between phonetic, orthographic, geographic and social factors.

## Algorithmic Translation of Names

As seen above, the mapping of names between two very different languages could be extremely complex. Although we have only illustrated such difficulty with Chinese and English, other language pairs might have the same problem.

But still, part of the solution might be programmable. For a system to be able to translate names algorithmically, naturally it needs to categorise and distinguish the various types of name in the first place, and then take the appropriate action correspondingly. We outline the major steps below.

Schematically, a system would first need to identify potential personal names from the source text (in Chinese). As mentioned in the beginning of this paper, there have been extensive studies on unknown word detection and name identification (e.g. Sun et al., 1995; Chen & Bai, 1998), and encouraging results have been reported.

Assuming that personal names could be satisfactorily identified from the source texts, the next step would be to recognise the origin of these potential names, i.e. whether a name is an authentic Chinese name or a translation. This is obviously a more difficult step than the first one. Nevertheless, the structure of names in terms of the number and combination of characters might give us some leverage. For instance, Sun et al. (2000) have shown different distributions of surnames and given names for names found from Beijing data and those from Hong Kong data.

The third and final step would be to find the proper renditions of the identified and categorised names in the target language (i.e. English). This might be done by looking up a large database of names (especially for the idiosyncratic cases) on the one hand, and possibly by some rule-based approach for the more regular cases, including the region-sensitive Romanisation, on the other.

In the following, we will study some existing MT systems with respect to their strategy and performance in personal name translation.

# Translation Systems

Four systems were evaluated in this study. All supported Chinese as a source language. Two were web-based, accessible over the Internet for free, and the other two were stand-alone commercial systems. Their main features are summarised below.

## EWGate

Developed by EWGate, Singapore, the system is available at http://www.EWGate.com/ewtranslite.html. It performs text translation of no more than 50 words each time. Bi-directional translation between English and Chinese (simplified / traditional), and that between English and Malay, are supported.

## WorldLingo

The system is developed by WorldLingo Inc., California, U.S.A. It performs text and URL translation, supporting ten source languages (English, French, German, Italian, Spanish, Portuguese, Chinese, Japanese, Korean and Russian) and all except Russian as target languages. The URL of the system is at:
http://www.worldlingo.com/products_services/worldlingo_translator.html

## TongYi (通譯漢英翻譯系統'98)

The system is developed by the Tianjin Datong Tongyi Software Research Institute, Tianjin, China. It supports Chinese-English document translation. In addition to a lexical database, it also keeps a sentence database. Source texts are pre-processed with word and sentence segmentation, and syntactic analysis, although no intermediate results are shown.

### Transtar V3.0 (譯星全文翻譯 3.0 版)

The system is developed by the CS & S Transtar Technology Company, China. It supports English-Chinese (E-C) and Chinese-English (C-E) document translation. According to the user manual, C-E translation is based on a vocabulary of over 100,000 words, and it claims 70% intelligibility and 200,000 words per hour for performance and speed respectively.

## Materials and Method

### The LIVAC Corpus and Lexical Database

LIVAC is a synchronous corpus developed by the City University of Hong Kong (Tsou et al., 2000). It consists of news media materials collected regularly and simultaneously since 1995 from six Chinese-speaking communities, namely Hong Kong, Singapore, Taiwan, Beijing, Shanghai, and Macau. The texts are automatically word-segmented with human verification. To date, the corpus has over 70 million Chinese characters with more than 400,000 words in its dictionary.

In addition to general vocabulary items, the LIVAC data also offer a rich source of proper names, including personal names, place names and organisation names. Hence the LIVAC database provides a leverage to assess the name coverage of the various MT systems.

### Test Data

The top 100 personal names according to media exposure in Hong Kong, Taiwan, and Beijing respectively, from 1999 to 2000, were used to test the various translation systems. Table 1 shows some example items from the three places. The aim is to study the mechanism of the translation systems and to assess the adequacy of the lexicons in them for the identification and translation of personal names.

| Hong Kong | Beijing | Taiwan |
|---|---|---|
| 陳水扁 (Chen Shui-bian) | 江澤民 (Jiang Zemin) | 陳水扁 (Chen Shui-bian) |
| 李登輝 (Lee Teng Hui) | 朱鎔基 (Zhu Rongji) | 連戰 (Lien Chan) |
| 宋楚瑜 (James Soong) | 李鵬 (Li Peng) | 宋楚瑜 (James Soong) |
| 江澤民 (Jiang Zemin) | 克林頓 (Bill Clinton) | 唐飛 (Tang Fei) |
| 成龍 (Jackie Chan) | 普京 (Vladimir Putin) | 蕭萬長 (Vincent Siew) |

Table 1 Personal Name Samples

Moreover, the most frequent personal names from different places exhibit different characteristics. They are representative of different social and linguistic contexts, and are hence useful for testing the robustness and orientation of the systems in translating proper names. The translation results were evaluated according to the following performance measures.

### Performance Measures

Each translation of the personal names was evaluated by a human judge according to a three-point scale:

0: Incorrectly translated / Not translated  
1: Partially correct translation  
2: Correct translation

Given the complexity of personal name translation between Chinese and English, we included the "partially correct" category to cover mainly the following two cases:

(a) the translation is not exactly in the conventionally accepted form, but either the last name or given name is correct, and

(b) the Romanisation does not follow the regional norm but nevertheless follows the standard Hanyu Pinyin.

## Results and Discussion

As said, the three places (Hong Kong, Beijing, and Taiwan) displayed quite different patterns of the most frequently found personal names in their news media. In the following, we first report the results obtained for each set of data individually, and then comment on the overall lesson learned.

### Hong Kong Data

Compared to Beijing and Taiwan, the Hong Kong index contains more names from entertainment and sports, where pop stars, soccer players, etc. receive so much attention and emphasis that gets them onto prominent positions of the index. Table 2 shows the performance of the four translation systems on the Hong Kong data.

| | EWGate | TongYi | Transtar | WorldLingo |
|---|---|---|---|---|
| Correct (C) | 24% | 5% | 9% | 15% |
| Partial (P) | 1% | 1% | 29% | 43% |
| C + P | 25% | 6% | 38% | 58% |

Table 2 Translation Accuracy for Hong Kong data

From Table 2, we see that on the whole WorldLingo managed to translate, at least partially, more than 50% of the top 100 names in Hong Kong. On the contrary, TongYi was the least satisfactory, about 10 times worse than WorldLingo and 4 to 5 times worse than the other two systems. On the other hand, although EWGate only came third in the overall accuracy, it did better than all others by the strictest criterion (C). About a quarter of the EWGate-translated names were perfect translations, whereas WorldLingo and Transtar only gave "partial" translations mostly. "Partial" translations, according to our criteria mentioned earlier on, could be a Hanyu Pinyin for a Chinese person instead of his or her actual English rendition (e.g. 曾蔭權, the Chief Secretary of Hong Kong, as "Zeng Yin Quan" instead of "Donald Tsang"; or 呂秀蓮, the Vice President of Taiwan, as "Lu Xiu Lian" instead of "Annette Lu"). It appears that WorldLingo and Transtar are relatively good at giving such Hanyu Pinyin for many Chinese names. However, under most circumstances, such a "secondary" rendition is of limited use and is not generally recognised by non-Chinese speakers.

As mentioned earlier, the Hong Kong data contain more people from entertainment. The results demonstrated that EWGate has an advantage over the others in this respect. Here are some examples of this kind which only EWGate got correct: 成龍 (Jackie Chan), 王菲 (Faye Wong), 陳慧琳 (Kelly Chen), and 張曼玉 (Maggie Cheung).

## Beijing Data

The personal names with top frequency in Beijing are mostly national and international political figures. Table 3 shows the performance of the four translation systems on the Beijing data.

| | EWGate | TongYi | Transtar | WorldLingo |
|---|---|---|---|---|
| Correct (C) | 30% | 6% | 20% | 56% |
| Partial (P) | 3% | 1% | 7% | 4% |
| C + P | 33% | 7% | 27% | 60% |

Table 3 Translation Accuracy for Beijing Data

For the Beijing data, WorldLingo is superior to the other three systems. Moreover, WorldLingo also came first even by the strictest criterion (C), unlike the Hong Kong case. EWGate, Transtar, and WorldLingo gave 30%, 20%, and 56% perfect translations respectively for Beijing names. TongYi is still the worst. Given that the Beijing samples have relatively more names from Mainland China, it is not surprising that WorldLingo, with its strength in Hanyu Pinyin (as discussed in the last section), has an advantage over the other systems. Furthermore, WorldLingo seems to have better coverage of the Chinese officials, such as 朱鎔基 (Zhu Rongji, Chinese Prime Minister), 李瑞環 (Li Ruihuan, Chinese People's Political Consultative Conference Standing Committee Chairman), 李嵐清 (Li Lanqing, Chinese Deputy Prime Minister), and 遲浩田 (Chi Haotian, Chinese Defense Minister).

Although WorldLingo performs better with Hanyu Pinyin names, in general EWGate seems to have a richer database for personal names, internationally speaking. For instance, only EWGate but not others correctly translated the following names: 普京 (Putin), 巴拉克 (Barak), 呂秀蓮 (Annette Lu, whereas WorldLingo only gave "Lu Xiu Lian"), 瓦希德 (Wahid), etc.

## Taiwan Data

The Taiwan index of personal names contains a considerable proportion of local, Taiwanese political figures, such as legislative members from different political parties. Table 4 shows the performance of the four translation systems on the Taiwan data.

| | EWGate | TongYi | Transtar | WorldLingo |
|---|---|---|---|---|
| Correct (C) | 19% | 0% | 5% | 16% |
| Partial (P) | 3% | 0% | 31% | 61% |
| C + P | 22% | 0% | 36% | 77% |

Table 4 Translation Accuracy for Taiwan Data

It was most striking that TongYi scored straight 0's. While EWGate performed fairly consistently as in the previous two sections, WorldLingo and Transtar got exceptionally many more "partial translations" than "complete translations" here. As mentioned, there are many local Taiwanese names in this sample set, and Taiwanese names often follow a Romanisation system different from Hanyu Pinyin. The strength of WorldLingo on Hanyu Pinyin is no longer an advantage. For instance, 李遠哲 (the President of Academia Sinica, Taiwan) was translated as "Li Yuan Zhe", while it should be "Lee Yuan-tseh"; and 張昭雄 (Deputy Chairman of People First Party, Taiwan) as "Zhang Zhao Xiong", while it should be "Chang Chao-hsiung", etc.

## Overall Comments

Generally speaking, the two web-based translation systems (i.e. EWGate and WorldLingo) performed better than the two stand-alone systems (i.e. TongYi and Transtar). TongYi, in particular, is extremely poor in this respect. It does not seem to have any mechanism to deal with personal name translation, but simply translates most names as a sequence of isolated and unconnected individual characters. Besides, apart from WorldLingo which also covered a variant Chinese rendition of Clinton from Taiwan, most other names from Taiwan were not rendered in their appropriate Romanised forms. Hence it is evident that none of the translation systems under evaluation caters for the Taiwan context.

Although EWGate is the best by the strictest criteria, it does not seem to have the ability to instantly put together some Hanyu Pinyin for possible names, unlike WorldLingo. This ability of WorldLingo not only makes it advantageous for the Beijing samples, but also provides some marginally acceptable solutions to many Chinese names in the absence of a standard entry in the lexicon.

It is hence obvious that different strategies are employed by different MT systems. The results above show that EWGate has a relatively comprehensive name database than other systems under evaluation, while WorldLingo has probably programmed the conversion of Chinese characters to Hanyu Pinyin. The results therefore suggest that a hybrid approach might be a possible solution for the complex task of personal name translation. On the one hand, we need large coverage of special transliterations, which could only be acquired from up-to-date corpora. On the other hand, where regularity is observed, we can go by algorithmic conversion.

## Conclusion

MT systems have to not only identify personal names from the source language but also render them in the appropriate forms in the target language. This depends to a large extent on the coverage of personal names in the lexical database of the systems. We have seen that systems varied considerably in this regard, and even the best system among those under test was far from satisfactory. The test results also suggest that a hybrid approach may be useful, such that part of the problem can be handled algorithmically and part of it by developing or acquiring a large database of cases which need special treatment. In fact, such a database would be useful not only in MT, but also in other applications like cross-language information retrieval. Given the complexity of name translation and its relation to the overall intelligibility of a translated text, the problem should by no means be overlooked in the future development of MT systems. Although we have only focussed on personal name translation between Chinese and English in this paper, the problem is extendable to other proper nouns as well as language pairs.

## Acknowledgements

This work was partially supported by the NTT Service Integration Laboratory and other organisations, but the findings and the views expressed thereon are the sole responsibilities of the authors.

## References

Arnold, D.J., Balkan, L., Meijer, S., Humphreys, R.L., and Sadler, L. (1994). *Machine Translation: an Introductory Guide.* London: Blackwells-NCC.

Chen K-J. and Bai, M-H. (1998) Unknown Word Detection for Chinese by a Corpus-based Learning Method. *International Journal of Computational Linguistics and Chinese Language Processing, 3(1):* 27-44.

EAGLES Evaluation Working Group (1999) *EAGLES Evaluation of Natural Language Processing Systems.* Final Report. EAGLES Document EAG-II-EWG-PR.1. Center for Sprogteknologi, Copenhagen.

Sun M., Huang, C., Gao, H. and Fang, J. (1995) Identifying Chinese Names in Unrestricted Texts. *Journal of Chinese Information Processing,* 9(2): 16-27.

Sun, M., Cheung, L. and Tsou, B.K. (2000) Finding Chinese Personal Names in Unrestricted Texts. Presented in the *Annual Conference and Joint Meetings of the Pacific Neighborhood Consortium (PNC 2000),* Hong Kong.

Tsou, B.K., Tsoi, W.F., Lai, T.B.Y., Hu, J. and Chan, S.W.K. (2000) LIVAC, A Chinese Synchronous Corpus, and Some Applications. In *Proceedings of the ICCLC International Conference on Chinese Language Computing,* Chicago, pages 233-238.