Corpus-based Extraction of Chinese Historical Term Translation Equivalents

Xiuying Li[1, 3], Chao Che[2], Xiaoxia Liu[1], Hongfei Lin[3], Rongpei Wang[1, 3]

1 School of Foreign Languages, Dalian University of Technology, Dalian, 116024, China

2 Key Laboratory of Advanced Design and Intelligent Computing, Dalian University, 116622, China

3 Dept. of Computer Science and Technology, Dalian University of Technology, Dalian, 116024, China

xyli2@126.com, chechao@gmail.com, hflin@dlut.edu.cn

Abstract

By closely examining the translation data of Chinese historical terms in some widely acclaimed English versions of Chinese historical classics, this paper has discovered a multi-equivalence phenomenon in term translation due to the polysemy of a source term and the synonymy of its potential equivalent concept in the target language as well as the different translation orientations by different translators in accordance with their considerations for the semantic and pragmatic dimensions of the translated term in its context for a particular readership. These findings exert a great challenge to the widely accepted basic assumptions underlying many statistical automatic terminological alignment models and the greedy algorithm applied to extract a valid translation pair from a bilingual parallel corpus. Therefore, this paper proposes to allow for the possibility of a multi-equivalent relationship between a source term and its potential target translation in the extraction algorithm design so that valid equivalent pairs would not be ruled out due to the exclusiveness of the algorithm.

Keywords

Term translation equivalent; corpus; Chinese historical term; English translation

1 Introduction

Term translation pair extraction is essential to cross-language information retrieval, natural language processing and machine translation. Considerable progress has been made in automatically aligning term translation pairs ever since the pioneering work of Gaussier, et al. (1992), Van der Eijk (1993), Daille et al. (1994), Dagan and Church (1997), Kwong et al. (2004) and Hippisley et al. (2005). However, the basic assumptions underlying these

statistical models for automatic terminological alignment are similar to the hypothesis initiated or applied for word-level alignment in the work of Brown et al. (1993), Melamed (1997) and Fung (1998):

1. Words have one sense per corpus
2. Words have single translation per corpus
3. No missing translations in the target document
4. Frequencies of bilingual word occurrences are comparable
5. Positions of bilingual word occurrences are comparable (Fung, 1998)

As Somers (2001) and Morin et al.(2007) point out, these simplified assumptions about the translation features of a source text and its target version do not always hold true. In the case of social sciences term translation, English translations of Chinese historical classics, in particular, the assumptions suffer from inherent weaknesses due to the contradiction between the hypothesis about the single sense and single best matching translation pair in one corpus generated on a one-to-one basis and the fact that, in many cases, there might be more than one valid English equivalent for a source Chinese historical term, as observed from the translation data of the bilingual parallel corpus of Chinese historical classics and their English translations. The multi-equivalence relationship between a source Chinese historical term and its candidate equivalents in the target language owes itself to the polysemy of a source term, the synonymy of its potential equivalent concept in the target language, and the different translation orientations by different translators in regard to the semantic and pragmatic dimensions of the translated term in its context for a particular readership. These findings make the competitive linking algorithm proposed by Melamed (1997) and the greedy algorithm model initiated by Gaussier et al. (1992) inadequate in matching the valid translation pairs in the bilingual parallel corpus of Chinese historical classics and their English translations, considering the fact that choosing statistically significant single best matching translation pair over other possible candidate pairs from a bilingual corpus based on a one-to-one assumption about the equivalent type would rule out potentially valid translation pairs in the corpus. Therefore, this paper proposes the extraction algorithm should allow for the possibility of a multi-equivalent relationship between a source term and its potential candidate translation so that valid equivalent pairs would not be ruled out due to the exclusiveness of the algorithm. The algorithm applied in this paper has achieved a recall score and a precision score of 85% and 91% respectively and F-measure 87.9 on a bilingual parallel corpus of the monumental Chinese historical classic *Shi Ji* (*Records of the Historian*) and its two English translations.

2 The Multi-equivalent Phenomenon of Chinese Historical Terms in English

A number of factors contribute to the multi-equivalence relationship between a Chinese historical term and its translation in English even in one corpus.

2.1 The Polysemy of a Chinese Historical Term

A Chinese historical term may be polysemous even in one corpus. For instance, the term "公" (*gong*) in *Shi Ji* may be used to refer to the title conferred by the ruler of the kingdom on nobles, as it was in the Zhou Dynasty in ancient China, in which the translation would

usually be “duke”, so “孝公” would be “Duke Xiao”. It may also be used to honor someone as a polite appellation, as in “楚南公”(*chunangong*) , which would be translated as “Master Nan of Chu” (Nienhauser, 2002), where “公” is translated into “Master”. As for someone who later gained important high social position, “公” would become “lord”, as in “汉王乃封侯公为平国君”(The king of Han enfeoffed Lord Hou as “Lord Who Pacifies the Nation” (Watson, 1961), which indicates that “公” is treated in the same way as the term “君”(*jun*, lord). For “公” in the term “沛公”(*peigong)*, which is translated into “governor of P’ei” -(Watson, 1961) and “Magistrate of P’ei” (Nienhauser, 2002), translators seem to have focused more on the semantic connotation of the term “公”, which was closer to “Magistrate” than “governor” and certainly in no way close to the social status of a “duke” or even “lord”, considering the context in which the title was given to indicate someone whose administrative duty was constrained to a county, though the historical figure of “沛公” did manage to make to the top of the power ladder by becoming the founding emperor of the Han Dynasty in ancient Chinese history. In the meantime, the term “公” may also be used as a polite addressing term for “you” in a conversation.

As for the term “三公” (*sangong*)，the evolution of its meaning is closely tied with the development of ancient Chinese history, too. When it appears in the documents about the Zhou Dynasty, it tends to be translated as “Three Dukes”, while in documents about other dynasties, it is often translated as “Three High Ministers or Officials” in regard to the different bureaucratic structures in different dynasties.

2.2 The Flexibility of a Chinese Historical Term in Terms of Part of Speech

A Chinese historical term may have multiple meanings because of its capacity for being used in a part of speech other than a noun, which inevitably results in diversified translations in English.

Take the term “卒” (*zu*) in *Shi Ji* for an example. It can be used as a noun, a verb and an adverb as well. When used as a noun, the term “卒” may refer to “soldiers” , “troops” or “foot soldiers”. When used as a verb, it may mean “to die or expire”. When used as an adverb, it may indicate “in the end” or “after all”. It may also become part of the phrase “士卒” (*shizu*), meaning “soldiers” , “troops” or “foot soldiers”.

Compared with “soldiers” and “troops”, “foot soldiers” sounds more archaic, but semantically it is more accurate. Similarly, translating the word “卒”, used as a verb, into “die” gives a more modern feel, which is more acceptable to the taste of contemporary readership than the alternative “expire”, although the latter retains the subtle sense of the source word, i.e. being an archaic verb for death.

To disambiguate the term “卒” from its usage of being a verb or an adverb, we might as well refer to the syntactic patterns exhibited clearly in its context. When used as a verb, “卒” is usually followed by a comma or a full stop and occurs together with another character “立” (*li*, setting up a new ruler). When used as an adverb, it is usually preceded by a punctuation of either a comma or a full stop and followed by other content words, mostly a verb. These contextual clues and the punctuation signs may be taken as significant factors in linguistically oriented term translation pair extraction and translation template extraction in natural language processing tasks.

Similarly, the term “王” (*wang*) does not mean only “king” in classic Chinese. It may be used as an addressing form, which might be translated as “Your Majesty” or simply as

"You" depending on the translator's preferences for the acceptable style in regard to its target readership. "Your Majesty" is formal and sounds a bit archaic stylistically in spite of being a close translation of the source term, which may satisfy the specialist's need for a concise meaning of the source term, while "You" may serve the need of the general readership as a colloquial everyday English addressing term. It may also be used as a pro-form indicating a ruler known to both the addresser and the addressee, which is often translated as "the king". Additionally, it may also be used as a verb, as in the phrase "王……(someone)", with the denotation of "set up someone as the king".

When used as a component of a phrase, "大王" (*dawang*), for example, it is not rendered as "king" any more. Instead, the two-character Chinese phrase is often taken as one addressing term, which is usually translated into "Your Majesty" or simply "You", subject to the preferred style of the translator in meeting the taste of his/her specific target audience.

2.3 The Synonymy of Potential English Equivalents for a Chinese Historical Term

Synonymy exists if two or more terms in a given language represent the same concept. Thus a synonym is a term used to designate the same concept as another term (Schmitz, 2006)

A Chinese historical term may designate a concept that has several equally valid equivalents in English. A good case in point is the high frequency term in *Shi Ji* "太子" (*taizi*, someone with the legitimate right to succeed to the throne when the reigning emperor passes away), which has been translated differently into English by renowned Sinologists and translators: namely, "heir apparent", crown prince (Watson, 1961), "Heir" (Nienhauser, 2002), "Heir Apparent" (Dubs, 1938), and "Crown Prince" (Bodde, 1940).

Results from the Cambridge Advanced Learner's Dictionary online (accessed on Mar. 10, 2010), shown in Table 1, seem to confirm the validity of these choices. The dictionary annotation and the examples provided indicate that "crown prince", "heir", "heir apparent" are all acceptable terms for referring to someone with the legitimate right to succeed to the throne, which could be regarded as positive evidence to demonstrate they are equivalents to the source term "太子".

crown prince	n. [C] the man who will be king of a country when the ruling king or queen dies.
Heir	n. [C]a person who will legally receive money, property or a title from another person, especially an older member of the same family, when that other person dies: e.g. The guest of honour was the Romanoff heir to the throne of all Russia.
heir apparent	n. [C, usually singular] the person with the automatic right to legally receive all or most of the money, property, titles, etc. from another person when they die: e.g. The Prince of Wales is the heir apparent to the throne.

Table 1. Dictionary entries of candidate terms for "太子"

Some Chinese historical terms are synonymous, too, which naturally leads to more English translation diversity. For example, the lexicalization forms for the concept of being

the most powerful ruling figure in a feudal society in Chinese would be "皇帝" (*huangdi*, emperor, August Emperor, a formal term), "皇上"(*huangshang*, emperor, ruler, sovereign, a title of respect), "上"(*shang*, emperor, ruler, sovereign, a colloquial expression), "万岁爷"(*wansuiye*, Lord of Ten Thousand years, emperor, a colloquial expression) and "圣上"(*shengshang*, emperor, sovereign, a title of respect), etc. The results from the Cambridge Advanced Learner's Dictionary online (accessed on Mar. 10, 2010) indicate that "sovereign" is synonymous with "ruler", "king" or "queen" in the sense of being the person with the highest power in a country. As an emperor is a ruler of a country, the word "emperor" is also synonymous with "sovereign".

2.4 The Different Translation Orientations by the Translator

The lack of an exact equivalent concept or a lexicalization of a similar concept in English culture and language often leads to different approaches to the translation of culturally loaded Chinese historical terms by even well-established translators or Sinologists. Their different orientations for the adequacy or the intelligibility of the translation or the taste of the potential target readership make the situation even more complicated.

Take, for instance, "左丞相"(*zuochengxiang*) and "右丞相"(*youchengxiang*) in *Shi Ji*. If translated literally, they would be "[C]chancellor of the left" and "[C]chancellor of the right" respectively, which do not make any sense to the general English readership as far as the power differences between the two posts are concerned. On the contrary, if translated sense by sense, they would be "Senior Chancellor" and "Junior Chancellor" respectively, as "左"(*zuo*) would be more important than "右"(*you*) in ancient Chinese culture. Similarly, for the term "氏" (*shi*), there would be three English candidates when referring to "family name": "cognomen, surname, family name"; two candidates when referring to "family": "clan, family". Here, "family name" and "family" are colloquial everyday English, while "cognomen and clan" sound more archaic and "surname" would lie stylistically in the middle. The translation decision-making about the specific choices would, to a great extent, depend on the translator's considerations for semantic and pragmatic dimensions of the translated term in its context for a particular readership.

These findings seem to pose a great challenge to the widely accepted basic assumptions underlying many statistical automatic terminological alignment models (Chang, 2003) and the greedy algorithm applied to extract a valid translation pair from a bilingual parallel corpus. It seems necessary to modify the widely accepted assumption for single one best-matching equivalent pair in each corpus if the insights gained from the previous research are to be implemented in extracting the historical term translation equivalents from a bilingual parallel corpus of Chinese historical classics and their translations. Therefore, this paper proposes to allow for the possibility of a multi-equivalent relationship between a source term and its potential target translation in the extraction algorithm design, and the final process of validating the equivalent status of the candidate target terms could draw from the data in an original English dictionary or such corpus as BNC (British National Corpus) or COCA (Corpus of Contemporary American English).

3. The Extraction Program

Most bilingual term translation pair extraction models begin with the identification of a candidate term in the source text based on observed features for termhood and then employ

the statistical algorithms to match its single best candidate equivalent in the target text (Somers, 2001). This approach has a basic problem when applied to classical Chinese text, which do not have clear word delimiters and would deem it necessary to segment the text into words before termhood computing is carried out. Yet there is no such a segmentation tool available for the general public, nor is there a publically available machine-readable general dictionary of classical Chinese. The lack of a bilingual dictionary of classical Chinese and English only makes the challenge even more insurmountable at the moment. To overcome these potential difficulties, the Chinese-English bilingual training corpus in this paper is sentence-aligned manually and a reference Chinese historical term list is prepared in consultation with a specialist in Chinese history and the word frequency list table in Li Bo (2006).

In order to maximize the extraction of more than one valid translation pair from the corpus, this paper proposes to set up a corpus of a Chinese historical text with each of its translations, as many of these texts have been retranslated, and then apply the program suggested below individually and retain the matches in each corpus before the final validation process is due.

The program in this paper is made up of two parts: the co-occurrence frequency approach and the head-word extension approach, the former mainly targeting at high-frequency term translation pairs while the latter at low-frequency ones.

Suppose the Chinese-English corpus is composed of {(CS_1, ES_1), (CS_2, ES_2),…, (CS_i, ES_i),…, (CS_n, ES_n)}, in which each group (CS_i, ES_i) represents an aligned sentence pair, with CS_i representing the Chinese sentence and ES_i the English sentence. To locate *ET*, the English translation equivalent of the Chinese term *CT*, this paper proposes the following steps:

(1) Check {t_1, t_2, …, t_i,…}, the serial number of the sentences which contain *CT* in the Chinese corpus, and count *F*, the frequency of *CT* in the corpus. Suppose each *CT* has been translated into an *ET*, then the sentence collection in which *CT*'s English equivalent *ET* appears would be ESS_t={ES_{t1} ,ES_{t2} ,…,ES_{tn}}.

(2) If *CT*'s frequency *F<3*, perform the head-word extension approach, and go to Step (4).

(3) If *CT*'s frequency *F>=3*, apply the co-occurrence approach, and go to Step (4).

(4) The extraction is over.

What follows is a detailed explanation of the two approaches involved in the term translation equivalent extraction process. As many terms are partially transliterated, the transliteration patterns have been integrated in both algorithms.

3.1 The Co-occurrence Approach

This approach is based on the following observations and assumptions: the historical terms in the Chinese corpus roughly co-occur with their equivalents in the English corpus as it is normally unlikely for a term to be omitted in historical classic translation. Therefore, we should be able to extract the term translation equivalents by checking the frequency count of the term in the Chinese corpus and that of its candidate equivalents in the English corpus. The candidate with an approximate frequency count would be the translation equivalent of the Chinese term.

Given the Chinese term *CT*'s English equivalent *ET*, the English sentence collection containing *ET* would be defined as ESS_t={ES_{t1} ,ES_{t2} ,…,ES_{tn}}, then

(1) Check the frequency count of each English word in the ESS_t, delete those that fit

$Fre(w_i) < \alpha * F$, in which w_i represents the English word, $Fre(w_i)$ the frequency count, and α the approximate coefficient, then we get the collection of English words that co-occur with *CT*, that is, $W=\{w_1, w_2, \ldots, w_i, \ldots\}$.

(2) If $\exists w_j \in ES_{ti}, w_j \in W$ and w_i occurs next to w_j within the sentences $ES_{ti} \in ESS_t$, and the co-occurring words $w_i \in ES_{ti}$ and $w_i \in W$, then link w_i and w_j as a *p* (phrase). If no co-occurring word appears next to w_i in ES_{ti}, w_i is taken independently as a *p*. Each sentence in ESS_t is then scanned so that the word collection *W* is turned into a co-occurring phrase collection $P=\{p_1, p_2, \ldots, p_i, \ldots\}$, in which p_i stands for the English phrase.

(3) Check each phrase p_i in the collection of *P*. If the phrase p_i only has one word and this word is also a stop word, then this phrase is deleted. Turn the first character in *CT* into a pinyin *py*. Check if *P* contains any phrase with a pinyin *py*. If yes, delete the p_i phrase without a *py*.

(4) Check the frequency count of each phrase in *P*, then choose the phrase with the highest count as a candidate translation equivalent of *CT*, that is, $ET = Arg_{p_i} Max(Fre(p_i)), p_i \in P$.

In view of the potential for one *CT* to be translated into more than one *ET*, the frequency count of the *CT* may not be strictly consistent with that of the candidate *ET*, we have adopted proportionality coefficient, that is, the ratio of the frequency of the *ET* in proportion to the *CT*, to improve the results. In the present experiment, if the frequency of the Chinese term F>6, the proportionality coefficient is α=2/3, otherwise α=1.

3.2 The Head-word Extension Approach

This approach mainly applies to the extraction of low-frequency Chinese term equivalents. The head-word within the term equivalent *ET* is located from the bilingual corpus before it is extended into a phrase through log likelihood ratio, which is calculated through the non-aligned corpus and functions as the criterion to determine whether two neighboring words are collocations. A bilingual sentence-aligned corpus and a large non-aligned English corpus are needed for the algorithm to work.

Given the Chinese historical term *CT* and the English sentence collection $ESS_t=\{ES_{t1}, ES_{t2}, \ldots, ES_{tn}\}$ in which the English translation equivalent *ET* appears, the head-word extension algorithm would work as follows:

(1) Transform each character in *CT* into its pinyin *py* and check if each sentence in ESS_t has a word in its *py* form. If yes, then add it to the head-word collection *HS*.

(2) If the head-word collection *HS* is empty, and *CT*'s frequency F>1, check $Fre(w_i)$, the frequency count of each English word in $ESS_{t,,}$ and then add w_i , the word with the highest frequency, into the head-word collection *HS*.

(3) If the head-word collection *HS* is empty, then the extraction fails. Otherwise, when the head-word is located, and the sentence containing this head-word ES_i is found, the neighboring words within the sentence context is obtained as $\{\ldots, w_{i-2}, w_{i-1}, hw_i, w_{i+1}, w_{i+2}, \ldots\}$.

(4) Judge if the neighboring words prior to the head-word within $\{\ldots, w_{i-2}, w_{i-1}, hw_i, w_{i+1}, w_{i+2}, \ldots\}$ can be extended into the head-word phrase *ET*. If w_{i-1} and hw_i are collocations, then add w_{i-1} into the head-word phrase *ET*. Then judge if w_{i-1} and its neighboring word w_{i-2} are collocations according to the log likelihood ratio results. Repeat this process until the neighboring word is judged to be unqualified to form a collocation or there is no more neighboring word found.

(5) Judge if the neighboring words which follow the head-word within {…,w_{i-2}, w_{i-1}, hw_i, w_{i+1}, w_{i+2},…}can be extended into the head-word phrase *ET*. If hw_i and w_{i+1} are collocations, then add w_{i+1} into the head word phrase *ET*. Then judge if w_{i+1} and its neighboring word w_{i+2} are collocations according to the log likelihood ratio results. Repeat this process until the neighboring word is judged to be unqualified to form a collocation or there is no more neighboring word found.

(6) The extraction is over.

To determine the collocation status of two words, we propose two hypotheses:

Hypothesis 1: suppose W_1 and W_2 are collocations;
Hypothesis 2: suppose W_1 and W_2 are not collocations.

Then the maximum log likelihood ratios of the two hypotheses are compared to decide whether W_1 and W_2 are collocations. Suppose c_1, c_2 and c_{12} represent the frequency count of w_1, w_2 and w_1w_2 in the corpus, w_1w_2 refers to the case when w_1 and w_2 co-occur, and N indicates the sum of the frequency counts of all the words in the corpus, the algorithm to calculate the collocation status of two neighboring words through the log likelihood ratio, adapted from Dunning (1993), would be like this:

$$\frac{L(H_1)}{L(H_2)} = \log L(c_{12}, c_1, p) + \log L(c_2 - c_{12}, N - c_1, p) - \log L(c_{12}, c_1, p_1) - \log L(c_2 - c_{12}, N - c_1, p_2) \quad (1)$$

Wherein, the parameters p, p_1 and p_2 in the formula are calculated as follows:

$$L(k,n,x)=x^k(1-x)^{n-k} \quad (2)$$

$$p = \frac{c_2}{N} \quad (3)$$

$$p_1 = \frac{c_{12}}{c_1} \quad (4)$$

$$p_2 = \frac{c_2 - c_{12}}{N - c_1} \quad (5)$$

4. The Extraction Experiment and Its Results

The training corpus used in this experiment is composed of five consecutive basic annals from *Shi Ji* and their corresponding English translations drawn respectively from *Records*

of the Grand Historian (Watson, 1961; 1993) and *The Grand Scribe's Records* (Nienhauser, 1994; 2002), both well-recognized authoritative translations, the former targeting mainly at the general English-speaking readership and the latter the Sinologists (Watson, 1961; Nienhauser, 1994). As the titles of the five basic annals are translated differently, Table 2 is presented to show the data source:

Shi Ji	Watson's Translation	Nienhauser's Translation
《秦本纪》	*Shih chi 5:Basic Annals of Qin*	*The Ch'in, Basic Annals 5*
《秦始皇本纪》	*Shih chi 6: The Basic Annals of the First Emperor of the Qin*	*The First Emperor of the Qin, Basic Annals 6*
《项羽本纪》	*Shih chi 7:The Basic Annals of Hsiang Yü*	*Hsiang Yü, Basic Annals 7*
《高祖本纪》	*Shih chi 8: The Basic Annals of Emperor Kao-tsu*	*The Exalted Ancestor, Basic Annals 8*
《吕太后本纪》	*Shih chi 9: The Basic Annals of Empress Lü*	*Empress Dowager Lü, Basic Annals 9*

Table 2. Experiment data source

There are 50,000 Chinese characters and 140,000 English words, including both English versions, in the non-aligned corpus. The same material is manually aligned into two bilingual parallel corpora of nearly 8000 Chinese-English sentence pairs or segments of sentence pairs. The reference list includes 362 Chinese terms. The following 3 formulas are adopted as the criterion in this experiment:

The precision score defined as *P*:

$$P = \frac{N_{correct}}{N_{translate}} \times 100\% \tag{6}$$

The recall score defined as *R*:

$$R = \frac{N_{correct}}{N_{all}} \times 100\% \tag{7}$$

The *F*-measure defined as disambiguation measurement:

$$F = \frac{2P \times R}{P + R} \tag{8}$$

Wherein, the $N_{correct}$, $N_{translate}$ and N_{all} stand for the number of the acceptable term translation pairs, the number of the candidate term translation pairs that have been extracted, and the total number of the term translation pairs in the whole corpus respectively.

If the extraction precision is considered to have reached its goal when results from either Watson's version or Neinhauser's version or both are acceptable, the recall score and the precision score amount to 85% and 91% respectively and the F-measure reaches 87.9. These results are much better than those in similar work done previously (Li et al., 2009).

The extraction results, presented in Table 3 and 4, also show the differences in term translation by the two translators, including the partially transliterated and the fully translated. It can be observed that the transliterations follow different systems, namely, the Mandarin Chinese *pinyin* (Column 2 in Table 3) and the Wade-Giles systems (Column 3 in Table 3). The former is gaining ground in China Studies in the West in recent years although the latter is still favored by many Sinologists even today.

Terms	Watson's Translation	Nienhauser's Translation
文公	Duke Wen	Duke Wen
竫公	Duke Jing	Duke Ching
夏桀	Xia Jie	Hsia Chieh
西戎	Western Rong	Western Jung
齐桓公	Duke Huan of Qi	Duke Huan of Ch'i
晋文公	Duke Wen of Jin	Duke Wen of Chin
周襄王	King Xiang of the Zhou dynasty	King Hsiang of Chou

Table 3. Transliterated titles

Terms	Watson's Translation	Nienhauser's Translation
高帝	Emperor Kao tsu	Kao ti
上将军	supreme general	Commander in Chief
御史大夫	imperial secretary	Grand Master of the Imperial Scribes
未央宫	Eternal Palace	Wei yang Palace
柱国	minister	Pillar of State
霸王	Dictator King	Hegemon

Table 4. Contrast of the translation approaches

5. Conclusion

We have presented two approaches to Chinese historical term translation pair extraction from a bilingual parallel corpus of Chinese historical classics and their English translations based on the observations concerning the English translation features of the Chinese historical terms, namely, the co-occurrence frequency approach and the head-word extension approach, targeting at high-frequency term translation pairs and low-frequency ones respectively. These approaches have performed well in the training corpus, considering the precision, recall scores and the *F*-measure achieved in the experiment and the extraction results, which seems to demonstrate the validity of the hypothesis proposed in this paper about the multi-equivalence relationship between a source Chinese historical term and its potential candidate English translations. It seems that the widely accepted assumptions, i.e. the single sense and single best matching translation pair in one corpus, in automatic term translation pair extraction algorithm are inadequate in Chinese historical term translation pair extraction and need to be modified to allow for the multi-equivalent pairs to emerge from the corpus so as to maximize the extraction effectiveness.

6. Acknowledgement

We are very grateful to the editors and the anonymous reviewers for their suggestions and insights for revision and the following grants for their support to this research: the 2009 Higher Institutions Research Project Fund of Liaoning Province of China (2009 A137), the 2010 Fundamental Research Fund for the Central Universities of China (DUT10RW201), the Natural Science Foundation Fund of China (No. 60673039, No. 60973068), National High Tech Research and Development Plan of China (No.2006AA01Z151) and the Doctoral Fund of Ministry of Education of China (No.20090041110002).

7. References

Bodde, D. 1940. Statesman, patriot, and general in ancient China: Three Shih Chi biographies of the Ch'in dynasty (255-206 B.C.). *Journal of the American Oriental Society*, 17: 1-75.

Brown, P. F., Pietra, S. A. D, Pietra, V. J. D. and Mercer. R. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2), 263-311.

Chang, B. 2003. Translation equivalent pairs extraction based on statistical measures. *Chinese Journal of Computers*, 26(5), 616-621.

Dagan, I. and Church, K. 1997. Termight: Coordinating humans and machines in bilingual terminology acquisition. *Machine Translation*, 12(1/2), 89–107.

Daille, B., Gaussier, É. and Lang ́e, J. M. 1994. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th Conference on Computational Linguistics*. Kyoto, Japan, pp. 515-521.

Dubs, H. H. 1938. *The History of the Former Han Dynasty: A Critical Translation with Annotations*. Waverley Press, Baltimore.

Dunning, T. 1993. Accurate method for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61-74.

Fung, P.1998. A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. In D. Farwell, L. Gerber and E. Hovy (eds.), *Machine Translation and the Information Soup*, Springer, Berlin, pp. 1–17.

Gaussier, E., Lang ́e, J. M. and Meunier, F. 1992. Towards bilingual terminology. In *Proceedings of the Joint ALLC/ACH Conference*. Oxford, pp.121-124.

Hippisley, A., Cheng, D. and Ahmad, K. 2005. The head-modifier principle and

multilingual term extraction. *Natural Language Engineering*, 11 (2), 129 – 157.

Kwong, O., TSOU, B., and LAI, T. 2004. Alignment and extraction of bilingual legal terminology from context profiles. *Terminology*, 10(1), 81-99.

Li, B. 2006. *Word Frequency Count of Shi Ji*. Beijing: Shangwu Press.

Li, X., Che, C., Han, L. and Liu, X. 2009. Extracting historical terms based on aligned Chinese-English parallel corpora. In *Proceedings of the 2009 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, Dalian, China, pp. 296-301.

Melamed, I. D. 1997. A word-to-word model of translational equivalence. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, Madrid, Spain, pp. 490-497.

Morin, E., Daille, B., Takeuchi K. and Kageura, K. 2007. Bilingual terminology mining – using brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp. 664–671.

Nienhauser, W. H. Jr.(eds.), Cheng, T. et al. trans. 1994. *The Grand Scribe's Records. Vol I: The Basic Annals of Pre-Han China*. Bloomington, Ind.: Indiana University Press.

Nienhauser, W. H. Jr.(eds.), Cao, W. et al. trans. 2002. *The Grand Scribe's Records, Vol II: The Basic Annals of Han China*. Bloomington, Ind.: Indiana University Press.

Schmitz, K.D. 2006. Terminology and Terminological Databases. In Brown, K. (eds.), *Encyclopedia of Language & Linguistics*. Elsevier Ltd., pp. 578-587.

Somers, H. 2001. Bilingual parallel corpora and language engineering. In *Proceedings of the Anglo-Indian Workshop "Language Engineering for South-Asian Languages" (LESAL)*, Mumbai, April, http://www.emille.lancs.ac.uk/lesal/somers.pdf, accessed on Feb. 30, 2008.

Van der Eijk, P. 1993. Automating the acquisition of bilingual terminology. In *Proceedings of the Sixth Conference on European Chapter of the Association for Computational Linguistics*. Morristown, NJ, USA, pp.113-119.

Watson, B. 1961. Records *of the Grand Historian of China Vol I & II*. New York: Columbia University Press. rev.: 1993. Records of the G rand Historian: Han Dynasty I & II. Hong Kong: Chinese University of Hong Kong Press.

Watson, B. 1993. *Records of the Grand Historian: Qin Dynasty*. Hong Kong: Chinese University of Hong Kong Press.