CHINET: a Chinese Name Finder System for Document Triage

K.L. Kwok, P. Deng*, N. Dinstl*, H.L. Sun*, W. Xu*, P. Peng§ and J. Doyon†*
*Computer Science Dept., Queens College, CUNY, 65-30 Kissena Blvd., Flushing, NY 11367
§Northrop-Grumman Information Technology
†MITRE Corp
kwok@ir.cs.qc.edu

Keywords: Foreign Language Processing, Information Search and Retrieval, Information Extraction

Abstract

CHINET (Chinese Name Extraction and Translation) is a prototype tool designed to assist analysts in performing Chinese document triage without being expert in the language. It is based on the premise that named entities are important information carriers, and that knowing their existence in a document will help analysts with proper intelligence analysis and exploitation of it. CHINET provides capabilities called tasks such as cross language name finding, name extraction and transliteration on single documents, pre-processed collections, or on web pages via Google access. Full scale translation of Chinese documents can also be provided by linking to facilities such as Systran's online machine translation capability. These tasks may enhance users' document triage operations and productivity. Results from extraction and transliteration evaluation shows that CHINET implements a useful concept for document triage.

1. Introduction

The searching, mining, analysis and management of information from the web or private text collections have become an important priority for many organizations with global interactions in this information-overloaded, web-based, multilingual society. Document triage - determining important, time-sensitive documents from unimportant ones - is shaping up as a necessary operation for achieving this function efficiently.

The objective of this project is to develop and implement a tool for analysts who have to monitor events about China via electronic Chinese texts, but who may not be experts in the language. They need to perform document triage on large amount of incoming Chinese documents and web pages. Our premise is that names of persons, places, organizations are strong carriers of information, and determining their presence in a document will greatly facilitate this task for analysts. Our tool is coined **CHINET**, an acronym for **CHI**nese **N**ame **E**xtraction and **T**ranslation. It is an interactive, GUI (graphical user interface)-based system using a client-server architecture. It employs the LINUX platform and can reside on a single laptop computer, or be accessed on a server via an IE browser.

This paper will describe CHINET's user capabilities and is organized as follows: Section 2 describes the basic functions implemented. Section 3 describes CHINET's capabilities available. Section 4 contains evaluation results, with conclusions in Section 5.

2. Basic Functions of CHINET

To achieve our objectives, CHINET provides four basic functions: 1) two-way translation of named entities between English and Chinese; 2) named entity extraction from Chinese documents; 3) name retrieval; and 4) general document indexing, management. 'Name' currently refers mainly to person and to some extent places. Location and organization names are also important, but at this stage our project limits the scope to mainly person names. These four functions are described below.

2.1 Bilingual Name Transliteration

Most statistical machine translation follows the IBM models (Brown, et al. 1990) as noisy channel transmission between source and target languages. The models require large amount of parallel bilingual training data in order to encode the observed regularities between two languages. For name transliteration however, these models appear less useful. There is in general less regularities in name formation and translation. Bilingual name list is not easily available, and new person and organization names are encountered or formed often and regularly. Name transliteration is recognized as a difficult task (Knight and Graehl 1997), yet critically important in contexts such as CLIR (cross language IR), CLQA (crosslingual question-answering) and MT (machine translation).

Our approach is to consider name transliteration as an information retrieval (IR) problem rather than a channel transmission loss problem by using the Web as a (noisy) 'bilingual dictionary' where name translations are embedded and to be mined (like translation memories). This is possible because of the immense scale, dynamic and timeliness nature of the WWW. Direct web retrieval of transliterations would be most successful for popular, current names. This is precisely the situation where bilingual lists or less-recently trained systems may suffer. Thus, web-based translation would complement nicely other types of MT.

Our idea is to exploit the normal convention for authors of a Chinese (or other language) document to make reference to a Western named entity via the patterns:

$S_C(S_E)$ or $S_E(S_C)$ (1a,b)

where S_C is a contiguous string of Chinese characters (delimited by a blank, punctuation, or ascii), and S_E is a string of English words within Chinese strings. They will have a high probability of containing embedded translation /transliteration of each other. These patterns have been exploited for other uses (Nagata, et al. 2001; Cao and Li 2002), but we may be one of the first to apply them for name transliterations (Deng and Kwok 2004). Recently (Lin, et al. 2004) has also reported excellent name transliteration results with these patterns.

Following are examples of such patterns that are found on the web:

1)… 丹尼尔卡纳曼(Daniel Kahneman)(以色列、美国)是第一位获得诺贝尔…

2)… 还有赛尚(Cezanne) 、雷诺瓦(Renoir) 、卢梭(Le Douanier Rousseau) 、…

3) … KONG FU MOVIE(功夫电影)、BRUCE LEE(李小龙)、JACKIE CHEN(成龙)。

4)… 敦豪环球速递公司(DHL)购入中…

5)… 首都: Monrovia(蒙罗维亚)官方语言为英语

6)…高密度脂蛋白(high density lipoprotein, HDL)各种脂蛋白的组成 …

7)… 北京车展大众汽车参展车型--高尔夫GTI(附图)

Examples 1)-3) show correct name transliterations found in the patterns. 3)-6) show that such patterns also provide correct translations for a company (DHL), location (Monrovia) and terminology (KONG FU MOVIE; high density lipoprotein, HDL). Example 7) is a case where the translation of GTI is wrong. It is much more often that the correct transliteration is present in such patterns. Moreover, further filtering operations may screen out such erroneous entries.

English-Chinese Name Transliteration

A named entity (in English) is used as a key K to search the web with request for Chinese GB-encoded snippets as output. String patterns satisfying (1a,b) are then identified. If none exists, web-assisted translation fails and other methods are used. The match between K and S_E is prioritized as exact ($K=S_E$), or partial ($K \subset S_E$). In texts S_E, a person name can be accompanied by titles such as 'Mrs.', 'President'. Partial match ($K \supset S_E$) is not considered because we break K into its components for searching if it is longer than one word and not found. Each pattern found is further divided into having case match or not. The string S_C that pairs with each S_E is segmented into substrings $s_C(i)$ from rightmost to the left recursively. Occurrence frequencies of these substrings f(i) are counted within the set of snippets. Each becomes a candidate of transliteration and is further filtered as follows: 1) if $s_C(i)$ occurs in different patterns, it is assigned to the best pattern priority; 2) if f(i) does not satisfy an occurrence threshold or not sufficient compared to max_of_f(i), $s_C(i)$ is deleted as noise; 3) if $s_C(i) \subset s_C(j)$, they are merged based on occurrence frequencies. Candidates are then ranked by pattern priority first, and within priority by length($s_C(i)$)*f(i). The top m (=5) candidates are considered as output. Often, with strict thresholds, less than five candidates would result.

This web-assisted translation is used to supplement two other resources: a verified bilingual person and place name list of ~20K, and an algorithmic transliteration from Pinyin names to Chinese characters (PYNAME for Chinese person names and GEONAME for Chinese place names (Kwok and Deng 2002, 2003)). Final outputs are ordered by bi-list matching first, candidates with highest votes, web, and algorithmic results. Matching existence on our bi-list is considered correct.

Chinese-English Name Transliteration

Almost any Chinese character, including name characters, can be mapped into Pinyin English output uniquely. Transliteration of Chinese named entities is therefore relatively simple. For back-transliterating a western person name, or when a Chinese person uses western first name (like James Soong), we also employ the above web-assisted translation procedure using the given Chinese name string as K, and still requesting Chinese-encoded output. This becomes less productive than the E-C direction, and often returns few snippets with useful patterns or none. A second round of web searching (as in pseudo-relevance feedback for IR) is done in order to improve the chance of locating useful patterns successfully. English words $w_E(j)$ from the first retrieval are captured, skip-words removed, ranked, and a new query: K AND ($w_E(1)$ OR .. $w_E(5)$) is formed using at most top five words. The detected English candidates are merged if there are substrings satisfying certain frequency thresholds. Final candidates are ranked by frequency.

2.2 Name Extraction from Chinese Texts

Chinese name extraction is difficult because, unlike English, there is no white space to delimit words or names. Word segmentation and name extraction can influence each other. This is an active research area: starting from rule-based approaches (Sun, et al. 1994), investigators have applied increasing sophisticated statistical methods such as Hidden Markov Model (HMM) (Sun, et al. 2003) and Support Vector classification (Isozaki and Kazawa 2002). The expert system approach has difficulty in prioritizing the rules and

their management when a system grows, but can be quite accurate in restricted domain applications. The automatic approaches need large amount of training data which is usually in short supply. CHINET makes use of both types, and a combination of their output to gain better effectiveness.

HMM-based Name Extraction

This system was implemented at Peking University and modified to interact with CHINET. A given Chinese sentence S_C may have many different word segmentation possibilities and result in different word path choices for the sentence. The best path satisfies: argmax $P(W|S_C) = P(S_C|W)*P(W)/P(S_C)$, where P(W) is the probability of word path W, with W including common words or names, and $P(S_C)$, $P(S_C|W)$ are constants and ignored. Generation of all possible paths W for a sentence depends on a large word dictionary and rules for generating Chinese and foreign names. In reality, there are far too many words to give good estimates for the probabilities, and instead, a more workable choice is the POS of each word. This leads to finding the best path that satisfies: argmax $P(T|S_C)= P(S_C|T)*P(T)$, where P(T) is the probability of a POS path that corresponds to W. POS information is available from the word dictionary. Bigram language model is used as an approximation to P(T). The component probabilities are trained from several months of hand-tagged newspapers. Viterbi decoding is employed to find the best POS path. Once a POS path is found, the word path can be identified together with the segmented words or names that are present.

Rule-based Name Extraction

For Chinese person names, we employed properties of name formation such as: sg, sg_1g_2, s_1s_2g, $s_1s_2g_1g_2$, where 's' and 'g' denote single characters used for surname or given name. (Names longer than 4 characters are possible but are very rare and ignored). We obtained probabilities of characters used in name formation from a monolingual name list of about ¾ million entries. Chinese surname characters are relatively closed (~600) and provide cues to locate candidate names in a document. Name probabilities are approximated as products of character probabilities trained from the monolingual list. They are also position dependent. Afterwards, a number of rules are used to give context weights (positive or negative) to the candidates. These include certain addressing title or job/position words (e.g. 先生|夫人|教授|总统|..) before or after a candidate, certain 'speaking verbs' (e.g. 说|回答|指出. .) located after a candidate that may characterize a person, organization or location cue words (e.g. 公司|集团..) after a name that may contradict its person name candidacy, etc. These are combined with the probability to arrive at a final score for each candidate.

For western names, it is customary to use a set F of 'foreign' characters (~400) to transliterate them to Chinese. In a document, the appearance of character sequences from F is therefore used as cues to detect foreign names. A probability is also evaluated for each name based on character or bigram probabilities trained from a set of foreign names. Context weights are also assigned as for Chinese names, and a final score is returned for each name. These Chinese and foreign name candidates are heuristically combined with the HMM-based candidates to provide a final output set.

2.3 Chinese Name Retrieval

Our PIRCS retrieval system (Kwok 2000) is employed to index name strings and Chinese documents/collection via consecutive overlapping bigrams as terms. Bigrams allow partial matches of the search name key when transliteration outputs have many forms or errors. The rank retrieved documents then go through a second round of ranking based on exact matches with the name strings. When multiple names are used as search keys, documents with more unique names will be ranked higher, followed by name occurrence frequencies. Currently we only output exact matches.

2.4 Document Processing, Management

To be a self-contained tool, CHINET needs to provide users with document processing and management functions. Commands are provided for users to load a document or collection from his/her own file folders to CHINET, and to map them to our default GB encoding when they are in BIG-5 codes. Documents are transformed to PIRCS's internal format and indexed with bigrams for retrieval. Collections can be named, stored or modified, and they will be reflected in the master GUI screen for searching and name extraction purposes.

3. CHINET Tasks

The basic functions described in Section 2 are combined to offer various capabilities for a user to perform document triage. We call these tasks: 1) English-Chinese cross language name finding (CLNF); 2) Name extraction, translation; 3) document and bilingual list management. The first two tasks can be applied to a single document by cut-and-paste, to pre-processed collections, or to web pages.

3.1 English-Chinese CLNF

A user may have a person name in mind and want to locate its occurrence in a document, collection or the Web. This is offered as 'CHINET Single' and 'CHINET Collection' buttons. The first option allows users to paste a document into a 'GetGBdocument' or 'GetBIG5document' box for searching. After one or more English name(s) are typed in the upper left input box, each is translated to candidate Chinese name sets (Sec.2.1). A user may uncheck any of these candidates if s/he knows that some are wrong. Users can also paste over the candidates. The 'Retrieve' function will locate and highlight all occurrences of the name candidates in the document displayed on the upper right box, and listed as bilingual pairs in a table in the lower right. Each table entry has a 'Trace' button that allows the user to step through each occurrence, which is useful if the document is long.

The second option allows a user to locate the name(s) of interest in pre-processed collections, or on the web via Google. Each document/page that has the name(s) is listed as a title in a table in a left lower box. By clicking on any document, its content will be displayed in the upper right with a table for 'Tracing' in the lower right as in the first option. CHINET also provides a link to online translation services (such as Systran) so that a retrieved Chinese web page can be displayed in English in full. MT services generally do not have good coverage for name transliteration.

For both options, CHINET also provides a 'GetNames' button which extracts, translates *all* names that occur in a displayed document (see Sec.3.2 below). This CLNF task therefore allows a user to locate a person name in the context of a document or a web page, and in companion with other names.

3.2 Named Entity Extraction

The main purpose of this task is for wholesale name extraction from a pre-processed Chinese collection. All possible name candidates are extracted and their occurrence frequencies captured. These are displayed in a bilingual table on the left middle screen and can be sorted by a user's choice: by frequency, by Chinese character order, etc. To avoid slow response when dealing with large collections, one can choose a smaller number of names to be translated or for display. When an extracted candidate name is clicked, documents containing it will be listed in another table below the previous one. From there, each document can be clicked and viewed in the right display window, or 'GetNames' to display all other names. This task can be extended to provide batch name-mining from a collection. The mined list can then be compared against a user-provided name list; this way name occurrence can be monitored in a batch fashion.

3.3 Document Management

This task provides users with capabilities to manage the documents, collections and bilingual name list efficiently. It includes: 1) collection/document loading, view or modification, and indexing. Indexing is needed to prepare a collection for searching purposes. 2) bilingual name list management. This allows a user to add/correct bilingual name pairs or delete wrong ones. These updates will eventually be reflected in improved accuracy and efficiency of the system.

4. Evaluation

After CHINET has been functioning, we attempt to evaluate its effectiveness for both name extraction and transliteration. These are described in the following subsections.

4.1 Name Extraction Evaluation

The focus of this evaluation portion of the project was to provide a baseline of the CHINET tool's extraction capability. For the development effort, Queen's College was tasked with building an analytic tool to identify, display, and transliterate both Chinese and Western person names within Chinese documents. The evaluation results of the CHINET system provide recall, precision, false positive, false negative, and F-measure scores based solely on these two named entity types.

The evaluation process included: corpus collection, ground truth development, CHINET evaluation, and results analysis. The corpus selected for the CHINET extraction evaluation was the Second Multilingual Entity Task Evaluation (MET-2) data (MUC 2001). The MET-2 (which ran in conjunction with the Seventh Message Understanding Conference [MUC-7]) was the follow-on to MET-1, an evaluation of Japanese, Chinese, and Spanish extraction systems that tagged named entities in newspaper articles (MET2 1997). The MET-2 Chinese data consists of 100+ documents with items tagged for all five of the MUC-defined named entities: 1) person, 2) organization, 3) location, 4) date/time, and 5) money/percentage. The corpus selected for this evaluation was 100 of the MET-2 Chinese documents. The ground truth was created from this same corpus by removing all tags that were not related to Chinese or Western person names, which resulted in a number of documents containing no tagged entities. In addition, all but 14 one-word names (e.g., 唐，陈: unlike "Cher" the single names here are abbreviated names, i.e., only last name, first name, or even part of the first name) were removed from the ground truth. Finally, the corpus was processed by the CHINET system and the output was compared and scored against the ground truth. The evaluation was scored using the metrics shown in Fig.1. Individual scores were calculated for each of the corpus documents for all five of the metrics listed. These scores were then averaged to produce overall system scores for the five metrics. They are listed in Fig.1.

Metrics	Definition	CHINET Result
a	number of correct that the system produced	230
b	number wrong that the system produced	220
c	number in ground truth that the system missed	18
Recall	R=a/(a+c)	92.74%
Precision	P=a/(a+b)	51.11%
False Positive	FP=b/(a+b)	48.89%
False Negative	FN=c/(a+b)	4.00%
F-measure	F=2PR/(P+R)	65.90%

Fig.1: CHINET Name Extraction Evaluation Results

Despite CHINET's occasional failure to identify person names correctly in context (e.g., it confuses location and organization names with person names), it shows promising utility in the following areas:

- Transliteration of English-spelled names into the corresponding Chinese characters; and

- Triage assistance for non-native Chinese analysts of softcopy Chinese documents by providing analysts the ability to search and locate person names within the Chinese text.

4.2 Name Transliteration Evaluation

As discussed before, our bilingual name transliteration resources consist of a bi-list of about 20K person and location entries that are considered correct. This is supplemented by web-based mining and algorithmic name Pinyin back-transliteration. To evaluate the transliteration accuracy of our approach, we need a verified bilingual name list that is openly available, objective and independent of our system (i.e. not used as part of our resources) to serve as ground truth standard. It should contain both person names of Chinese origin and non-Chinese origin. Unfortunately we could not locate one, and are using Wikipedia's translated person name list 'elist' (Wiki2005a) and its Chinese person name list 'clist' (Wiki2005b) instead.

'elist' provides a set of 363 raw English name entries with their Chinese translation(s). These are of western origin and style. Most have multiple words (e.g. given name, middle, family name), and their Chinese counterpart are separated by a GB-encoded 'dot'. They were broken into single word entries because web documents usually do not contain full name translations. This breakup creates a file of 493 unique single English entries (with some non-name words like 'homo', 'romance', etc. removed), and they total to 692 unique pairs. We manually determined that there are 406 person names, 74 western and 1 Chinese place names and 12 others. This becomes one of our ground truth standards called 'wike-ec'. By switching the role of English with Chinese, we formed another 'standard' called 'wike-ce' that will be used to evaluate CHINET for transliterating western names in Chinese back to English. The following show some example entries:

Antisthenes 安提斯泰尼		Aquinas 阿奎拿	亚奎拿
Boffrand 鲍夫朗		Diocletian 戴克里安	
Stewart 史都华	司徒华	Zwickau 茨维考	
Thomas 多马	多马斯田纳西	托玛斯	托马斯

'clist' provides another set of 465 unique GB-encoded Chinese person names with no translation. These are semi-automatically mapped with dictionary consultation to Pinyin (sometimes multiple) to form Chinese-English pairs. This is possible because *mapping characters to Pinyin is almost always unique, but the reverse is not true*. Examples are:

曾子 Ceng Zi, Zeng Zi 梅兰芳 Mei Lan Fang
曹丕 Cao Pi 温家宝 Wen Jia Bao, Wen Jie Bao

By using each English Pinyin as key and corresponding Chinese entry as answer, we set up a third 'standard' called 'wikc-ec'. It is used to evaluate CHINET for back-transliterating Chinese person names in English to Chinese. Again, by reversing the role of English and Chinese we defined the fourth 'standard' 'wikc-ce' for evaluating Chinese to English transliteration of Chinese names. This seems superfluous in view of the italics above. However, this test will use CHINET's automatic transliteration procedures. For example, when a Chinese sequence is given to CHINET, it does not know if it represents a Chinese person or place, and various steps of translation may suggest different candidates other than the standard Pinyin.

These Wikipedia name lists are not limited to newspapers, web or current affairs domains (from where CHINET is trained). It has ancient person names (like Antisthenes, Marcianus, 曾子, 曹丕) and others that may not be currently popular on the web. This represents a severe challenge to CHINET's web-based approach, and evaluation results should be viewed within this context. If evaluated using names of more current nature, we believe CHINET would provide higher effectiveness.

Each unique English (Chinese) name in a 'standard' is used as input to CHINET and the top 10 rank-ordered Chinese (English) outputs are compared to the answer(s). If there is *exact match*, a 1 is assigned to that rank, otherwise 0. Some names have multiple 1's at different ranks because of multiple answers. Table 2 shows % correct at different ranks 1-6. Column 's3' (meaning 'success-in-3') gives percent of names with at least 1 correct answer within top 3 and will be used to characterize effectiveness. s10 is also shown.

Rank>	**1**	**2**	**3**	**4**	**5**	**6**	**s10**	**s3**
English-to-Chinese: 493 Western Names								
wike-ec	44.2	17.4	3.0	2.4	.4	.4	65.1	**62.3**
wike-ec+	46.9	18.3	3.4	2.6	.4	.4	68.4	**65.5**
English-to-Chinese: 479 Chinese Names								
wikc-ec	60.3	11.3	3.1	2.1	1.0	1.7	82.7	**74.5**
Chinese-to-English: 639 Western Names								
wike-ce	34.4	6.3	.9	.5	.2	.0	41.8	**41.2**
Chinese-to-English: 660 Western Names								
wike-ce+	35.6	6.5	.9	.5	.2	.0	43.2	**42.6**
Chinese-to-English: 465 Chinese Names								
wikc-ce	90.3	10.3	.9	.6	.2	.2	99.4	**98.7**

Table 2: CHINET Name Transliteration Evaluation Results

First row 'wike-ec' shows that, for this test of western person names from anywhere in the world, 44.2% is transliterated to Chinese characters correctly at the first suggestion by CHINET. Rank 2 correct is 17.4%. After scanning top 10 candidates, 65.1% of input names have at least 1 correct transliteration (s10). For top 3 only, s3 equals 62.3%.

Often, a western name transliterated to Chinese differs from the ground truth by 1 or 2 similar sounding characters (see 'Stewart', 'Thomas' above). Our evaluation considers it wrong because it has no exact match, but it is not necessary so. For those 34.9% (1-.651) names that have no correct answers, we further consult another objective translation name list by (About 2005), as well as output from (Systran 2005) translation of the English as alternates. Row 'wike-ec+' shows that these additional Chinese 'ground truth' entries boost the result by ~5% to s3=65.5%. Back-transliterating Chinese person names from English is an easier task as shown in 'wikc-ec' where s3=74.5%. We do not use augmented ground truth for these. For a mixture of

western and Chinese person name inputs, we would expect s3~70%. This English-Chinese (E-C) transliteration capability helps in the CLNF task, and is a good aid for analysts to find transliterations.

Back-transliterating from Chinese to English for western names (wike-ce+) is much harder: rank1 correct is only ~ 35%, and s3 ~ 42%. For Chinese person names however (wikc-ce), it is very accurate as expected: s3=98.7%.

5. Conclusion and Discussions

CHINET is a name-based prototype Chinese document triage system. Its major capabilities are name extraction, translation and cross language name finding. We combine statistical and rule-based name extraction methods that evaluate MET2 data to an F-measure ~66%. Currently, our project concentrates on person names only. Our novel web-based approach to the difficult task of name transliteration provides good accuracy for Chinese person names: E-C ~75% and C-E ~99% success in suggesting correct transliterations within top 3 ranks. For western names, effectiveness is E-C ~65%, C-E ~42%. This is evaluated against a name list of general nature, not just newspaper/current affairs oriented, and may be a lower bound to CHINET's effectiveness. CHINET can also serve as a model for prototypes of other languages.

CHINET's name extraction can be enhanced by adding focus on location and organization name detection, plus more syntactic analysis to help identify context more accurately. These will improve extraction precision by not classifying these names as person. Name transliteration can be improved by using more sophisticated patterns or exploiting web page structures. CLNF can also be made more sophisticated during searching by adding contexts for specialized domains. CHINET has been set up at http://xpkk.cs.qc.edu:8080/chinet/ for users to try its capabilities.

Acknowledgments

This material is based upon work funded in whole or in part by the U.S. Government and any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

We would like to thank Rich Hall who made valuable suggestions for our GUI design, and the Government COTR for his support throughout this project.

References

(About 2005) http://chineseculture.about.com/library/name/blname.htm

Brown, P, Cocke, J, Della Pietra, S, Della Pietra, V, Jelinek, F, Lafferty, F, Mercer, R & Roossin, P, 1990. A statistical approach to machine translation. *Computational Linguistics* 16(2):79-85.

Cao, Y & Li H. 2002. Base noun phrase translation using web data and the EM algorithm. *COLING-2002*, pp.127-133.

Deng, P & Kwok, K.L. 2004. A cross language name finding system. In: First Intl. Joint Conf. on NLP, IJCNLP-04, Interactive Posters/Demos, pp.9-12.

Gao, J.F, Nagata, M, Saito, T & Suzuki, K. 2001. Using the web as a bilingual dictionary. In: Proc. of ACL 2001 Data Driven-MT Workshop. pp95-102.

Isozaki, H. & Kazawa, H. 2002. Efficient support vector classifiers for named entity recognition. COLING 2002, paper C02-1054.

Knight, K. and Graehl, J. 1997 *Machine transliteration.* Proc.of 35th Annual Meeting of ACL, pp. 128-135.

Kwok, K.L. 2000. "Improving English and Chinese Ad-Hoc Retrieval: A Tipster Text Phase 3 Project Report". Information Retrieval, 3:313-338.

Kwok, K.L. & Deng, P. 2002. Corpus-based Pinyin Name Resolution. Proc. 1st SIGHAN Workshop on Chinese Language Processing (COLING 2002). pp. 41-47

Kwok, K.L & Deng, Q. 2003. GeoName: a system for back-transliterating pinyin place names. Proc. HLT-NAACL 2003 Workshop: Analysis of Geographic References. pp.26-30

Lin, T, Wu, J-C & Chang, J.S. 2004. Extraction of name and transliteration in monolingual and parallel corpora. In: AMTA 2004 Proceedings. pp.177-186. Lecture Notes in Artificial Intelligence 3265: Springer.

(MET2 1997) http://torvald.aksis.uib.no/corpora/1997-2/0038.html

(MUC 2001) http://www.itl.nist.gov/iaui/894.02/related_projects/muc/muc_data/muc_data_index.html

Nagata, M, Saito, T & Suzuki, K. 2001. Using the web as a bilingual dictionary. In: Proc. of ACL 2001 Data Driven-MT Workshop. pp.95-102.

Sun, J, Zhou, M and Gao, J. 2003. A class-based language model approach to Chinese named entity identification. Computational Linguistics and Chinese Language Processing. 8(2):1-28.

Sun, M.S, Huang, C.N, Gao, H.Y & Fang, J. 1994. Identifying Chinese names in unrestricted texts. Communications of COLIPS. 4(2) pp. 113-122.

(Systran 2005) http://www.systranbox.com/systran/box

(Wiki 2005a) http://zh.wikipedia.org/w/index.php?title=%E8%AF%91%E5%90%8D%E8%A1%A8/A&varian

(Wiki 2005b) http://zh.wikipedia.org/wiki/%E4%B8%AD%E5%9B%BD%E4%BA%BA%E5%88%97%E8%A1%A8